Using Next-Generation Sequencing To Understand the Aetiology of Dystonia and Other Neurological Diseases

A thesis submitted to the University College London

for the degree of Doctor of Philosophy

August 2015

By

Dr Gavin Charlesworth

Declaration of Authorship

I, Gavin Charlesworth, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis presents my work using both next-generation and traditional genetic techniques aimed at further clarifying the aetiology of hereditary neurological disorders, with a particular focus on dystonia. A large part of this was focused on the identification, clinical phenotyping, and genetic analysis of kindreds with neurological disease inherited in a Mendelian fashion but for which no causal mutation had yet been identified. My work led directly to the discovery of two new dystonia genes, *ANO3* and *HPCA*, and the identification of two novel phenotypes for the known disease-associated genes, *ATM* and *NUBPL*. Mutations in a third novel gene, *SLC25A46*, was identified as the most likely cause of disease in a kindred with a complex neurological disorder consisting of optic atrophy, severe action myoclonus and peripheral neuropathy, but could not be confirmed due to lack of a second segregating kindred – it was published, a year and a half after we had first identified it, by another group, just as I was in the process of submitting thesis. Taken together, these results confirm that whole exome sequencing in combination with linkage analysis or homozygosity mapping represents a powerful means of dissecting out the genetic aetiology of Mendelian disease.

In addition, this thesis summarises my foray in the world of association analyses, a technique which underpins the recent explosion in knowledge regarding the genetic architecture of so-called 'complex' disease. I use this technique to show that the association signal over *MAPT* in Parkinson's disease survives when affectation status is defined neuropathologically. Finally, I present my work using traditional Sanger sequencing to better understand the prevalence of already published Mendelian disease genes for both dystonia and Parkinson's disease.

Table of Contents

Acknowledgements

I would like to begin by thanking my supervisors, Professors Nicholas Wood and John Hardy, for offering me the opportunity to complete my PhD in a laboratory at the forefront of genetic research. Their guidance, support, encouragement and advice have been essential not only to the completion of my research itself, but also to the much more arduous process of putting it down on paper in the form of this thesis. There were times when it seemed I would never get there. I also owe a debt of gratitude to Professor Kailash Bhatia who identified many of the families that I worked with during my thesis and demonstrated his faith in my ability to deliver the goods by passing their details to me.

I would also like to express my thanks to the 'pure scientists' with whom I had the pleasure of working during this time. Their friendliness and willingness to help, in the face of my often-astounding ignorance of even basic scientific processes, was really quite heart-warming. I leave my time in research with a real admiration for these people, whose working days are long and often extend into the weekend and whose efforts do not receive the true financial and social recognition that I believe they ought to. Although I cannot name everyone, I feel that I must nonetheless single out the following individuals for special thanks: Dr Andrey Abramov, Dr Kira Holmström, Dr Plamena Rumenova, Dr Fernando Bartolomé-Robledo, Dr Mina Ryten, Dr Daniah Trabzuni and Dr Elisavet Preza. These people not merely offered me friendliness and general advice, but gave time from their own extremely busy schedules to help me complete essential aspects of my work that I would have been incapable of completing alone. Many of my best publications would not have been possible without their help.

I consider myself very privileged during my time in research to have become particularly good friends with some very amazing people. Una Sheerin, Alan Pitman, Rob Wykes, Niccolo Mencacci and Deborah Hughes – your down-to-earth and occasionally rather drunk and dirty approach to life kept a smile on my face during those three years and ensured that what I most remember about my PhD is the laughs rather than the hard work.

On a personal note, I would like to thank my parents for their unwavering support of whatever I have chosen to do, their downright Northern and unpretentious enjoyment of life, and their frank and sincere empathy for others: these things serve as a constant yardstick of goodness for me. To my Dad, for working so hard throughout his life to ensure that I could have opportunities he did not and for being a great enough person to seek to create a different life for himself and his children to that which he had known himself; to my Mam, who is at once no-nonsense, but full of love and infectiously engaged with the world around her: I hope that I can make you both proud in what I do.

No man is an island, or so said John Donne, and I couldn't have completed this work without the support of people whose main contribution was not in the production of the science but in the keeping of me sane. I owe a debt of gratitude to all my friends in this respect, but there are a few I would particularly like to single out. To Diego: *estuviste a mi lado y me ayudaste a volver como debe ser. No te lo puedo agradecer lo sufficiente. Siempre serás alguien muy especial para mí.* To Clem: *on a commencé ce voyage ensemble et, meme si l'on ne l'a finit pas ainsi, je veux néanmoins te remercier de ton appui, de ta gentillesse, de ton amitié, et de ton amour – toujours.* To Angus: your support has been invaluable. Without you, I would undoubtedly be shipwrecked. There is no way I can say thank you, except simply to say thank you. I really wish you and all your family all the best.

List of Publications

First Author Publications:

Mutations in *HPCA* cause autosomal-recessive primary isolated dystonia. Charlesworth G, Angelova PR, Bartolomé-Robledo F, Ryten M, Trabzuni D, Stamelou M, Abramov AY, Bhatia KP, Wood NW. Am J Hum Genet. 2015 Apr 2;96(4):657-65. doi: 10.1016/j.ajhg.2015.02.007. Epub 2015 Mar 19.

No Pathogenic *GNAL* Mutations in 192 Sporadic and Familial Cases of Cervical Dystonia. Charlesworth G, Bhatia KP, Wood NW. Mov Disord. 2014 Jan;29(1):154-5. doi: 10.1002/mds.25713. Epub 2013 Nov 12.

Ataxia Telangiectasia Presenting as Dopa-Responsive Cervical Dystonia. Charlesworth G, Mohire MD, Schneider SA, Stamelou M, Wood NW, Bhatia KP. Neurology. 2013 Sep 24;81(13):1148-51. doi: 10.1212/WNL.0b013e3182a55fa2. Epub 2013 Aug 14.

The Genetics of Dystonia: New Twists in an Old Tale. Charlesworth G, Bhatia KP, Wood NW. Brain. 2013 Jul;136(Pt 7):2017-37. doi: 10.1093/brain/awt138. Epub 2013 Jun 17.

Primary and Secondary Dystonic Syndromes: an Update. Charlesworth G, Bhatia KP. Curr Opin Neurol. 2013 Aug;26(4):406-12. doi: 10.1097/WCO.0b013e3283633696.

Mutations in *ANO3* cause dominant craniocervical dystonia: ion channel implicated in pathogenesis. Charlesworth G, Plagnol V, Holmström KM, Bras J, Sheerin UM, Preza E, Rubio-Agusti I, Ryten M, Schneider SA, Stamelou M, Trabzuni D, Abramov AY, Bhatia KP, Wood NW. Am J Hum Genet. 2012 Dec 7;91(6):1041-50. doi: 10.1016/j.ajhg.2012.10.024. Epub 2012 Nov 29.

Tau Acts as an Independent Genetic Risk Factor in Pathologically Proven PD. Charlesworth G, Gandhi S, Bras JM, Barker RA, Burn DJ, Chinnery PF, Gentleman SM, Guerreiro R, Hardy J, Holton JL, Lees A, Morrison K, Sheerin UM, Williams N, Morris H, Revesz T, Wood NW. Neurobiol Aging. 2012 Apr;33(4):838.e7-11. doi: 10.1016/j.neurobiolaging.2011.11.001. Epub 2012 Jan 4.

Screening for *VPS35* Mutations in Parkinson's Disease. Sheerin UM*, Charlesworth G*, Bras J, Guerreiro R, Bhatia K, Foltynie T, Limousin P, Silveira-Moriyama L, Lees A, Wood N. Neurobiol Aging. 2012 Apr;33(4):838.e1-5. doi: 10.1016/j.neurobiolaging.2011.10.032. Epub 2011 Dec 7.

Acute, Localised Paroxysmal Pain as the Initial Manifestation of Focal Seizures: a Case Report and a Brief Review of the Literature. Charlesworth G, Soryal I, Smith S, Sisodiya SM. Pain. 2009 Feb;141(3):300-5. doi: 10.1016/j.pain.2008.11.005.

Non-First Author Publications:

Analysis of Parkinson's Diease Brain-derived DNA for Alpha-Synuclein Coding Somatic Mutations. Proukakis C, Shoaee M, Morris J, Brier T, Kara E, Sheerin UM, Charlesworth G, Tolosa E, Houlden H, Wood NW, Schapira AH. Mov Disord. 2014 Jul;29(8):1060-4. doi: 10.1002/mds.25883. Epub 2014 Apr 21.

The phenotypic spectrum of DYT24 due to *ANO3* mutations. Stamelou M, Charlesworth G, Cordivari C, Schneider SA, Kägi G, Sheerin UM, Rubio-Agusti I, Batla A, Houlden H, Wood NW, Bhatia KP. Mov Disord. 2014 Jun;29(7):928-34. doi: 10.1002/mds.25802. Epub 2014 Jan 17.

C9ORF72 Expansions, Parkinsonism, and Parkinson disease: a Clinicopathologic study. Cooper-Knock J, Frolov A, Highley JR, Charlesworth G, Kirby J, Milano A, Hartley J, Ince PG, McDermott CJ, Lashley T, Revesz T, Shaw PJ, Wood NW, Bandmann O. Neurology. 2013 Aug 27;81(9):808-11. doi: 10.1212/WNL.0b013e3182a2cc38. Epub 2013 Jul 24.

Migraine with Aura as the Predominant Phenotype in a Family with a *PRRT2* mutation. Sheerin UM, Stamelou M, Charlesworth G, Shiner T, Spacey S, Valente EM, Wood NW, Bhatia KP. J Neurol. 2013 Feb;260(2):656-60. doi: 10.1007/s00415-012-6747-4. Epub 2012 Nov 24.

Study of the Genetic Variability in a Parkinson's Disease Gene: *EIF4G1*. Tucci A, Charlesworth G, Sheerin UM, Plagnol V, Wood NW, Hardy J. Neurosci Lett. 2012 Jun 14;518(1):19-22. doi: 10.1016/j.neulet.2012.04.033. Epub 2012 Apr 23.

Consortium-Based Publications:

Large-scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, DeStefano AL, Kara E, Bras J, Sharma M, Schulte C, Keller MF, Arepalli S, Letson C, Edsall C, Stefansson H, Liu X, Pliner H, Lee JH, Cheng R; International Parkinson's Disease Genomics Consortium (IPDGC); Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI); 23andMe; GenePD; NeuroGenetics Research Consortium (NGRC); Hussman Institute of Human Genomics (HIHG); Ashkenazi Jewish Dataset Investigator; Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE); North American Brain Expression Consortium (NABEC); United Kingdom Brain Expression Consortium (UKBEC); Greek Parkinson's Disease Consortium; Alzheimer Genetic Analysis Group, Ikram MA, Ioannidis JP, Hadjigeorgiou GM, Bis JC, Martinez M, Perlmutter JS, Goate A, Marder K, Fiske B, Sutherland M, Xiromerisiou G, Myers RH, Clark LN, Stefansson K, Hardy JA, Heutink P, Chen H, Wood NW, Houlden H, Payami H, Brice A, Scott WK, Gasser T, Bertram L, Eriksson N, Foroud T, Singleton AB. Nat Genet. 2014 Sep;46(9):989-93. doi: 10.1038/ng.3043. Epub 2014 Jul 27.

Unbiased Screen for Interactors of Leucine-Rich Repeat Kinase 2 Supports a Common Pathway for Sporadic and Familial Parkinson Disease. Beilina A, Rudenko IN, Kaganovich A, Civiero L, Chau H, Kalia SK, Kalia LV, Lobbestael E, Chia R, Ndukwe

K, Ding J, Nalls MA; International Parkinson's Disease Genomics Consortium; North American Brain Expression Consortium, Olszewski M, Hauser DN, Kumaran R, Lozano AM, Baekelandt V, Greene LE, Taymans JM, Greggio E, Cookson MR. Proc Natl Acad Sci U S A. 2014 Feb 18;111(7):2626-31. doi: 10.1073/pnas.1318306111. Epub 2014 Feb 7.

Susceptibility Loci for Pigmentation and Melanoma in Relation to Parkinson's Disease. Dong J, Gao J, Nalls M, Gao X, Huang X, Han J, Singleton AB, Chen H; International Parkinson's Disease Genomics Consortium (IPDGC). Neurobiol Aging. 2014 Jun;35(6):1512.e5-10. doi: 10.1016/j.neurobiolaging.2013.12.020. Epub 2013 Dec 27.

Genetic Comorbidities in Parkinson's Disease. Nalls MA, Saad M, Noyce AJ, Keller MF, Schrag A, Bestwick JP, Traynor BJ, Gibbs JR, Hernandez DG, Cookson MR, Morris HR, Williams N, Gasser T, Heutink P, Wood N, Hardy J, Martinez M, Singleton AB; International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2); North American Brain Expression Consortium (NABEC); United Kingdom Brain Expression Consortium (UKBEC). Hum Mol Genet. 2014 Feb 1;23(3):831-41. doi: 10.1093/hmg/ddt465. Epub 2013 Sep 20.

Analysis of Genome-Wide Association Studies of Alzheimer Disease and of Parkinson Disease to Determine if these 2 Diseases Share a Common Genetic Risk. Moskvina V, Harold D, Russo G, Vedernikov A, Sharma M, Saad M, Holmans P, Bras JM, Bettella F, Keller MF, Nicolaou N, Simón-Sánchez J, Gibbs JR, Schulte C, Durr A, Guerreiro R, Hernandez D, Brice A, Stefánsson H, Majamaa K, Gasser T, Heutink P, Wood N, Martinez M, Singleton AB, Nalls MA, Hardy J, Owen MJ, O'Donovan MC, Williams J, Morris HR, Williams NM; IPDGC and GERAD Investigators. JAMA Neurol. 2013 Oct;70(10):1268-76.

Serum Iron Levels and the Risk of Parkinson Disease: a Mendelian Randomization Study. Pichler I, Del Greco M F, Gögele M, Lill CM, Bertram L, Do CB, Eriksson N, Foroud T, Myers RH; PD GWAS Consortium, Nalls M, Keller MF; International

Parkinson's Disease Genomics Consortium; Wellcome Trust Case Control Consortium 2, Benyamin B, Whitfield JB; Genetics of Iron Status Consortium, Pramstaller PP, Hicks AA, Thompson JR, Minelli C. PLoS Med. 2013;10(6):e1001462. doi: 10.1371/journal.pmed.1001462. Epub 2013 Jun 4.

The Val158Met COMT Polymorphism is a Modifier of the Age at Onset in Parkinson's Disease with a Sexual Dimorphism. Klebe S, Golmard JL, Nalls MA, Saad M, Singleton AB, Bras JM, Hardy J, Simon-Sanchez J, Heutink P, Kuhlenbäumer G, Charfi R, Klein C, Hagenah J, Gasser T, Wurster I, Lesage S, Lorenz D, Deuschl G, Durif F, Pollak P, Damier P, Tison F, Durr A, Amouyel P, Lambert JC, Tzourio C, Maubaret C, Charbonnier-Beaupel F, Tahiri K, Vidailhet M, Martinez M, Brice A, Corvol JC; French Parkinson's Disease Genetics Study Group; International Parkinson's Disease Genomics Consortium (IPDGC). J Neurol Neurosurg Psychiatry. 2013 Jun;84(6):666-73. doi: 10.1136/jnnp-2012-304475. Epub 2013 Feb 13.

A Pathway-Based Analysis Provides Additional Support for an Immune-Related Genetic Susceptibility to Parkinson's Disease. Holmans P, Moskvina V, Jones L, Sharma M; International Parkinson's Disease Genomics Consortium, Vedernikov A, Buchel F, Saad M, Bras JM, Bettella F, Nicolaou N, Simón-Sánchez J, Mittag F, Gibbs JR, Schulte C, Durr A, Guerreiro R, Hernandez D, Brice A, Stefánsson H, Majamaa K, Gasser T, Heutink P, Wood NW, Martinez M, Singleton AB, Nalls MA, Hardy J, Morris HR, Williams NM. Hum Mol Genet. 2013 Mar 1;22(5):1039-49. doi: 10.1093/hmg/dds492. Epub 2012 Dec 7.

Using Genome-Wide Complex Trait Analysis to Quantify 'Missing Heritability' in Parkinson's Disease. Keller MF, Saad M, Bras J, Bettella F, Nicolaou N, Simón-Sánchez J, Mittag F, Büchel F, Sharma M, Gibbs JR, Schulte C, Moskvina V, Durr A, Holmans P, Kilarski LL, Guerreiro R, Hernandez DG, Brice A, Ylikotila P, Stefánsson H, Majamaa K, Morris HR, Williams N, Gasser T, Heutink P, Wood NW, Hardy J, Martinez M, Singleton AB, Nalls MA; International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2).

Hum Mol Genet. 2012 Nov 15;21(22):4996-5009. doi: 10.1093/hmg/dds335. Epub 2012 Aug 13

Use of Support Vector Machines for Disease Risk Prediction in Genome-wide Association Studies: Concerns and Opportunities. Mittag F, Büchel F, Saad M, Jahn A, Schulte C, Bochdanovits Z, Simón-Sánchez J, Nalls MA, Keller M, Hernandez DG, Gibbs JR, Lesage S, Brice A, Heutink P, Martinez M, Wood NW, Hardy J, Singleton AB, Zell A, Gasser T, Sharma M; International Parkinson's Disease Genomics Consortium. Hum Mutat. 2012 Dec;33(12):1708-18. doi: 10.1002/humu.22161. Epub 2012 Aug 3.

A Two-Stage Meta-Analysis Identifies Several New Loci for Parkinson's Disease. International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2). PLoS Genet. 2011 Jun;7(6):e1002142. doi: 10.1371/journal.pgen.1002142. Epub 2011 Jun 30.

Imputation of Sequence Variants for Identification of Genetic Risks for Parkinson's Disease: a Meta-Analysis of Genome-Wide Association Studies. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW. Lancet. 2011 Feb 19;377(9766):641-9. doi: 10.1016/S0140-6736(10)62345-8. Epub 2011 Feb 1.

List of Figures

List of Tables

CHAPTER 1:

Introduction

1. Introduction

1.1 Genetics and Neurological Disease

As the control centre of the body, the human nervous system represents one of the most complex biological structures known. It mediates not only our ability to interact with the environment via sensation and movement, but is also the seat of those higher faculties – cognition, memory, speech, thought and emotion – that combine to give each human being their particular individual nature. The correct development and maintenance of this exquisitely intricate structure require the co-ordinated action of a large number of genes, often with precise spatial and temporal patterns of expression. Moreover, although the nervous system exhibits considerable plasticity – a prerequisite to environmental adaptation and learning – this is evident predominantly at the synaptic level. In contrast, on a cellular level, nervous tissue is characterised by a relatively limited capacity for regeneration once damaged. Thus, when considered in the context of the relatively long life span of a human being, genetic variation that results merely in 'inefficiencies' or moderate dysfunction of the pathways involved in cellular maintenance may alone be sufficient to cause progressive neuronal dysfunction and, ultimately, cellular death, manifesting clinically as neurodegenerative disease. Taken together, these facts go some way towards explaining the vulnerability of the nervous system to genetic disease. Even considering monogenic forms of genetic disease alone, it has been estimated that at least 60 - 70% involve the nervous system [1]. Many such Mendelian disorders – motor neuron disease, Alzheimer's disease, Parkinson's disease and dystonia to name but a few – have so-called 'sporadic' counterparts that affect a much larger number of individuals and thus represent a substantial burden on patients and carers, as well as a major challenge in terms of healthcare provision and its associated costs for society as a whole. A significant proportion of these disorders show increasing prevalence with age, such that the problem is becoming increasingly acute as life expectancy climbs, the elderly population grows in size worldwide, and the cost of caring for affected individuals threatens rise to unsustainable levels. In 2010, in an attempt to tackle this challenge, 22 EU countries launched the first Joint Programming Initiative to combat neurodegenerative disease,

based on a shared vision to speed up the progress towards new treatment options, to identify preventative strategies, and to improve the care of those already affected.

A prerequisite step in attempting to treat or even prevent any neurological disorder is gaining an understanding of what exactly has gone wrong. An understanding of the pathogenesis of human disease is thus a key goal for researchers working within the health sciences. Armed with knowledge of the cellular and molecular events that underpin a particular condition, clinicians and scientists can begin to develop truly targeted therapies, designed either to interrupt the underlying biological cascade at some key nodal point or, alternatively, to modulate critical cellular pathways that are dysfunctional. However, deciding which molecules and pathways are of aetiogenetic importance for a particular disease, from amidst the myriad that make up the exquisitely complex machinery of the cell, is an almost Herculean task. What is needed is a 'way in' – a protein, robustly connected to the disease, that can serve as a starting point for the further exploration of the wider molecular biology of the condition. At present, one of the best available means of identifying such a protein in relation to any particular disease remains the discovery of causal genes linked to monogenic forms of the disorder. Knowledge of a gene causal for a Mendelian disorder provides the identity of a protein whose dysfunction is alone sufficient to cause disease. If the function of this protein is known or can be experimentally established, then it can be placed within a cellular pathway that is presumably important for pathogenesis. From there, cellular and animal models can be created that permit the testing of therapies designed to prevent, to slow, or – at the very least – to compensate to some degree for the disease process, with the hope that some may show sufficient promise to make it to human trials and, possibly, beyond. Furthermore, it has long been recognised that disease-modifying agents aimed at neurodegenerative disorders are most likely to be effective if they are implemented at an early point in the course of the disease, ideally at a preclinical stage. At present, identification of presymptomatic individuals carrying high-risk Mendelian mutations represents one of the most promising means of assembling such cohorts for future clinical trials, thus further reinforcing the importance of genetics throughout the process of developing successful treatment strategies for disease.

Mendelian forms of neurological diseases represent, however, only a tiny proportion of the clinical disease burden. Why expend so much time and effort elucidating the genetic cause of such rare forms of disease? In fact, it has long been hypothesised that the insights gained from genetic analysis of Mendelian disorders will be relevant – to some degree at least – to the more common, sporadic forms of these disorders. In sporadic forms of disease, it is envisioned that disease may result from mild to moderate 'ineffeciencies' in at least some of the same cellular pathways responsible for Mendelian disease, acting in combination with other individual genetic and environmental factors. Recently, the results of genome wide association studies (see section 1.7 below) – undertaken in relatively common sporadic neurodegenerative disorders, such as Alzheimer's or Parkinson's disease – have provided the first solid evidence to support this belief: common variants in some of the same genes that are known to cause Mendelian forms of a disorder also modulate the risk of developing sporadic disease [2-5].

1.2 Genetics and Mendelian Disease

Three decades or so prior to my own work in this area, the process of genetic analysis of Mendelian disease began in earnest. Using the genetic techniques available to them at the time – linkage analysis, positional cloning and, subsequently, candidate gene studies – and often only after painstaking years of work, researchers were able to gradually flesh out the list of known causal genes for Mendelian forms of many hereditary conditions. Neurology has been one of the leaders in applying the developing tools of genetics to understand the aetiology of disease. Extending as far back as the early 1980s, these new methods were used to identify first the location of the Huntington disease gene, Alzheimer disease genes, and the first Parkinson disease gene. Extremely rare, highly penetrant mutations in these genes were subsequently shown to be causative for disease. These discoveries redirected entire fields of study and greatly improved our understanding of the underlying pathophysiology of these conditions.

The identification of the genes responsible for Mendelian disease was critically dependent on the identification of large, well-characterised pedigrees, groups of pedigrees with a presumed identical genetic aetiology, or consanguineous families with

disease most likely resulting from homozygous mutations. This was not always a simple matter given the relative rarity of Mendelian disease as well as the tendency for smaller family sizes and falling rates of consanguinity in much of the world today (particularly that part of the world where the technology existed to study such families). The critical importance of phenotypic characterisation to the success of linkage analysis also meant that issues such as phenotypic heterogeneity, reduced penetrance or late-onset disease (where the affectation status of the youngest generation cannot be determined) made some diseases inherently difficult to tackle. Beyond the relative scarcity of suitable pedigrees, other practical factors severely limited progress, in particular the huge cost and physical time commitments required to sequence the large number of candidate genes that were covered by linkage peaks. As a result – and though nonetheless an impressive feat – it is estimated that the genetic basis of less than 50% of all Mendelian disorders had been determined by 2010 [6], when I started my research. Experience from the specialised neurogenetics clinics of the National Hospital for Neurology and Neurosurgery echoes this estimation – there remained numerous kindreds exhibiting ostensibly Mendelian disease, but for which no causal gene could be identified. A significant part of this thesis focuses on my attempts to solve the puzzle of the cause of disease in these kindreds by application of the most recently developed genetic sequencing technologies, in combination with more traditional linkage and homozygosity mapping strategies. In the following sections, I describe the early impact and potential of this new genetic technology, whilst discussion of its technical aspects and associated methodologies can be found in chapter 2 and 3.

1.3 Next-Generation Sequencing and Its Impact

DNA sequencing was first developed in 1975 by Sanger and Coulson [7]. Over the next three and a half decades, gradual improvements in technology and reagents led to a streamlining of the process, but the underlying chemistry – which remained fundamentally unchanged – imposed an upper limit on the speed and scale of the process. The arrival of next-generation sequencing (NGS) technologies in the late 2000s allowed researchers to break through this glass ceiling. Unlike Sanger sequencing, which is based on the electrophoretic separation of chain-termination products generated by individual sequencing reactions, next-generation sequencing

(NGS) involves the massively parallel sequencing of clonally-amplified or single-DNA molecules that are spatially separated on a flow cell. Sequencing is performed by repeated cycles of polymerase-mediated nucleotide extensions. As a massively parallel process, NGS generates hundreds of gigabases of nucleotide-sequence output in a single instrument run and has the capacity to sequence the whole genome in days to weeks.

Despite the fact that an individual laboratory could now feasibly sequence an individual's whole genome, this remained – particularly in the early years after the advent of NGS – a relatively costly endeavour. Moreover, from the point of view of the identification of causal mutations in genetic disease, ascertaining the likely functional consequences of variants in non-coding parts of the genome is extremely difficult, given the relatively poor understanding of what these regions (which were, until recently, branded as 'junk') actually do. Therefore, more commonly, an adaptation of the same basic technique is used to sequence selected regions from across the genome that are presumed to be of higher importance. As around 85% of all disease-causing mutations identified (prior to the advent of next-generation sequencing) were exonic [8, 9], sequencing strategies that targeted only the protein-coding regions of the genome (the 'exome') represented an attractive compromise. Accounting for only ~1% of the genome, the exome can be sequenced at a much lower cost and in a much quicker time, whilst still generating sequence data for those regions in which a causal mutation is most likely to be found. This strategy, known as whole exome sequencing (WES), is that which is most commonly employed in current attempts to identify causal genes for unexplained Mendelian disease and is also that which I used in my own efforts.

1.4 The Challenge of Human Genetic Variability

The advent of NGS technologies permitted large scale projects aimed at sequencing the exomes of numerous individuals, including large cohorts of ostensibly healthy people of various ethnicities. Such sequencing projects have revealed that human beings tolerate a surprisingly high degree of genetic variation.

With reference to disease in particular, two important insights have been gained. Firstly, there is a relative excess of low frequency variants most likely resulting from the

recent demographic explosion in the human population [10-14]. This event has allowed the accumulation in the human genome of an excess of new mutations and rare alleles, skewing the pattern of genetic variation. Furthermore, such low frequency variants are enriched for nucleotide changes that affect protein function, particularly those that are predicted to be deleterious, and are therefore more likely to be related to disease [10-15]. Given their recent origin, the persistence of such variants may be explained to some extent by the lack of sufficient evolutionary time for their removal by natural selection. Nonetheless, some degree of purifying selection has taken place as evidenced by the observation that nonsynonymous variants are significantly under-represented compared with synonymous variants [15-17]. Secondly, most rare variants are restricted to particular populations and show very little mixing between the continents [13, 14, 18, 19].

In fact, it is now known that ostensibly healthy individuals can – and indeed do – carry many apparently disadvantageous variants without exhibiting any obvious ill effects. This situation can arise for a number of reasons – healthy individuals may: 1) carry only a single disease allele for a recessive disorder; 2) carry a disease-causing variant for a disorder that has its onset late in life and may not reach the required age to manifest; 3) carry a disease-causing variant for a disorder that requires additional environmental or genetic factors to manifest (i.e. reduced penetrance); or 4) carry a variant resulting in a clinical phenotype that is mild enough be classified with the range of normal healthy variation (i.e. reduced potential intelligence).

Experience to date suggests that analysis of an average exome would be expected to reveal between approximately 20,000 to 24,000 single nucleotide variants, depending on the population from which the DNA sample originates (see table 1 overleaf) [20]. The vast majority of these variants will have been recorded in publically available databases of genetic variation and thus will most likely (but not necessarily) be non-disease causing polymorphisms. However, between 300 to 500 per exome variants are 'novel', i.e. never previously recorded in such databases. Of these, around two thirds will alter coding and about 1 – 2% will either produce a premature stop codon or affect a splice site. Novel missense, nonsense and splice site variants are associated with a greater *a priori* probability of being deleterious and thus, even considering these categories of

variants alone, the average exome will contain approximately 200 – 300 variants that could potentially cause disease [20]. A more recent study examined sequencing data from individuals and classified variants as damaging (on the basis of a Condel score† of >0.99) or disease-causing (on the basis of a known causal association with a Mendelian disease). On the basis of their study, the authors estimate that the average individual will carry >400 damaging variants and approximately 5 disease-causing mutations [21].

These facts present obvious challenges for the use of WES data in experiments designed to dissect out the genetic architecture of both sporadic and Mendelian disease. As this is not the main focus of my thesis, the difficulties such revelations pose in relation to dissecting out the genetic architecture of sporadic disease are touched on only briefly in section 1.9. In terms of Mendelian disease, however, the main problem centres around how to recognise a true, disease-causing variant from amongst a dense background of genetic variation that includes not merely polymorphisms but also hundreds of potentially damaging alleles. To date, researchers have employed a number of different strategies to overcome this difficulty, with choice depending on a number of factors, including the nature of the phenotype under consideration, the mode of inheritance, pedigree structure and likelihood of genetic heterogeneity. As these have direct relevance to the work presented in this thesis, the sections that follow provide an overview of how some of these approaches work as well as some of their possible pitfalls and limitations.

† Condel is a method to assess the outcome of non-synonymous SNVs using a CONsensus DELeteriousness score that combines various tools (MutationAssessor and Functional Analysis through Hidden Markov Models).

† Generally dbSNP135 or below, as more recent versions are flooded with variants from NGS projects involving individuals with disease phenotypes or projects assessing somatic variation.

† The work presented in this chapter was peer-reviewed and published with adaptations in Charlesworth *et al.*, 2013 [108].

Table 1 - Typical number and distribution of variants observed by exome sequencing in two different populations. This table was originally published in Bamshad *et al.*, 2011 [20].

Variant Type	Mean Number of Variants (± sd) in African Americans	Mean Number of Variants (± sd) in European Americans
Novel Variants		
Missense	303 (± 32)	192 (± 19)
Nonsense	5 (± 2)	5 (± 2)
Synonymous	209 (± 26)	109 (± 16)
Splice	2 (± 1)	2 (± 1)
Total	520 (± 53)	307 (± 33)
Non-novel variants		
Missense	10,828 (± 342)	9,319 (± 233)
Nonsense	98 (± 8)	86 (± 6)
Synonymous	12,567 (± 416)	10,536 (± 280)
Splice	36 (± 4)	32 (± 3)
Total	23,529 (± 751)	19,976 (± 505)
Total Variants		
Missense	11,131 (± 364)	9,511 (± 244)
Nonsense	103 (± 8)	93 (± 6)
Synonymous	12,776 (± 434)	10,645 (± 286)
Splice	38 (± 5)	34 (± 4)
Total	24,049 (± 791)	20,283 (± 523)

1.5 Outline of the Strategies Used for Identifying Causal Variants in Mendelian Disease

Strategies for honing in on the true causal variant for a Mendelian disorder from amongst the numerous variants in a typical exome sequencing dataset can be broadly divided in two: restricting the genomic search space and variant prioritisation. In order to narrow down the areas of the genome where the causal variant must lie (or, at the very least, is more likely to lie), either linkage analysis or homozygosity mapping can be employed. A fuller explanation of these methods can be found in sections 2.9 and 2.10, respectively. However, it is worth pointing out here that these techniques, particularly linkage analysis, are entirely reliant on accurate designation of affectation status within the pedigree. In particular, erroneous assignment of unaffected status to individuals that that are not thought to exhibit the disease phenotype but do, in fact, carry the causative mutation (either because the are non-manifesting carriers of a mutation that is not fully penetrant or because there is variable expressivity and the phenotype they manifest is not recognised as a possible manifestation of the same disease) has the potential to significantly reduce or even abolish linkage over the critical area. This is especially likely to happen if linkage analysis is performed under a model that assumes a too high a penetrance or if the pedigree is small. The scientific literature is littered with examples of published linkage loci in which no causal variant could be identified. Some of these have subsequently been shown to be incorrect when the kindred is eventually sequenced for a newly-identified disease gene and is found to carry a mutation in that gene; re-examination of the family in the light of the sequencing data often reveals sample handling errors, critical non-manifesting carriers or phenocopies (see, for example, Wider *et al.* [2008] [22] or Weber *et al.* [2011] [23]).

Even with the benefit of accurate linkage or homozygoisty data to narrow down the genomic areas under consideration, there would usually still be far too many undifferentiated variants to ever allow a researcher to easily identify the true causal variant within a particular kindred. In order to overcome this problem, additional data for each variant, drawn from publically available databases, must be sought and associated with the variant in a process called annotation. The lowest level of annotation, which is, to a large extent, completed automatically by freely available

scripts such as Annovar, involves the collation of a basic 'character profile' of each variant. Such a character profile typically includes: 1) the variant's consequence in terms of transcription and translation (e.g. coding, non-coding or spice site, synonymous or non-synonymous, frameshift or non-frameshift, ect.); 2) its previously reported frequency in disease free controls; 3) the level of pre-computed evolutionary conservation at and immediately surrounding its genomic position; and 3) its predicted pathogenicity according to various commonly used *in silico* softwares. Of these characteristics, the observed frequency of the variant in databases of normal human sequence variation and its effect on protein transcription are the two most commonly used means to rapidly reduce the number of variants under consideration. However, even here, there are potential pitfalls, as detailed below.

1.5.1 Rarity as a Means of Variant Filtration

Mendelian disease is rare and the genetic mutations that underlie it are generally highly penetrant. As such, one would not, as a rule, expect to find a variant that is truely causal for a Mendelian disorder in publically-available databases of normal human sequence variation. The most common databases used for this purpose are earlier version of dbSNP[†], the 1000 Genomes project, the NHLBI Exome Sequencing Project, and the Complete Genomics 69 databases. In-house databases of exome sequenced in individuals without the phenotype under question can also be used for this purpose. Unfortunately, experience teaches that some of the variants that are contained within these databases of normal sequence variation will subsequently be shown to be causal for Mendelian disease. For example, the common p.G2019S mutation in *LRRK2*, which is causal for a autosomal dominant form of Parkinson's disease, was initially present in dbSNP. The most common reasons for the appearance of Mendelian variants in such databases are much the same as those enumerated previously in the discussion of why healthy individuals may habour deleterious variants: reduced penetrance or age-related penetrance, meaning that non-manifesting carriers or individuals who will develop disease only at a later date may be sequenced and labelled as healthy; or, alternatively, heterozygosity for a highly penetrant recessive allele. This

[†] Generally dbSNP135 or below, as more recent versions are flooded with variants from NGS projects involving individuals with disease phenotypes or projects assessing somatic variation.

potential for 'contamination' means that filtering variants against these databases independently of their minor allele frequency entails a risk of accidentally eliminating a true causal variant. Instead, it is standard and accepted practice to eliminate variants from consideration if they are reported in such databases at a particular frequency or greater (usually determined by the presumed inheritance pattern and frequency of the phenotype under consideration). In general, for the genetic analysis of rare dominant and recessive Mendelian disorders, filtering variants on a maximum minor allele frequency of 0.1% and 1%, respectively is considered to be effective at reducing the number of variants under consideration without significantly reducing the power of the study to detect the causal variant [20].

Another consideration arises from the fact that many rare variants are private to particular populations (see section 1.3 above) [13, 14, 18, 19]. Therefore, some neutral rare variants that appear completely novel with respect to databases of sequence variation in one population (and would thus be considered more likely to be disease-causing) may, in fact, be seen in healthy individuals in another population. This emphasises the need for population-specific databases of sequence variation. Unfortunately, such databases simply do not exist for all populations. For various reasons, this is principally true of populations that tend to have higher levels of consanguinity and is thus a particular concern for studies of autosomal recessive disease thought to result from biallelic mutations inherited as a consequence of autozygosity.

1.5.2 Using Kindred Structure as a Means of Variant Filtration

When there is some idea of the likely mode of inheritance of a disease, kindred information is also frequently used to reduce the number of variants under consideration. As already mentioned, disease-causing variants in Mendelian disease are usually rare. For very rare alleles, the probability of identity-by-descent given identity-by-state is high even among distantly related individuals. For example, two first cousins share a rare allele that is identical-by-descent in approximately one-eighth of the genome. For dominant kindreds, sequencing the two most distantly related individuals with the phenotype of interest and then considering only those variants that are shared between both individuals can substantially restrict the genomic search space [20]. The

main difficulty lies in identifying kindreds with affected individuals who are sufficiently separated genetically, notwithstanding with the small, but non-negligible risk that one of the individuals chosen for sequencing may turn out to be a phenocopy and has therefore been deemed to be affected incorrectly.

In kindreds with likely autosomal recessive inheritance but without evidence of consanguinity, filtering for homozygous or compound heterozygous mutations, especially when combined with linkage data, can be a powerful method for honing in on the causal variant(s) [24]. Where there is evidence of consanguinity, on the other hand, the search can usually be restricted (initially at least) purely to homozygous variants lying in regions of autozygoisty, as determined by homozygoisty mapping based on genome-wide genotyping data, that are shared exclusively by all affected individuals,. In these cases, usually only the DNA of a single affected individual need be exome sequenced – essentially, the data is merely being used as a cost-effective, time-saving replacement for Sanger sequencing of all the genes in a pre-defined, crucial interval. Nonetheless, at least three caveats need to be borne in mind when tackling kindreds with ostensibly autosomal recessive disease. Firstly, an affectation pattern suggestive of an autosomal recessive mode of inheritance can occasionally be produced by dominant mutations when neither parent is affected, most often as a result of germline mosaicism or reduced penetrance. Secondly, in a small, but not insignificant number of consanguineous kindreds with autosomal recessive inheritance of disease, the causative mutations will, in fact, turn out to be compound heterozygous rather than homozygous mutations [25]. Thirdly, and of particular relevance if the disease phenotype is not uncommon in a particular population, locus heterogeneity and/or allelic heterogeneity existing within the same family can confound attempts to delineate a shared haplotype or region of shared autozygosity, as has been demonstrated in several studies [26-29].

Although not of direct relevance to the work presented in this thesis, disease secondary to *de novo* mutations can be tackled by exome sequencing of multiple parent–child trios. As multiple *de novo* events occurring within a specific gene (or within a gene family or pathway) can reasonably be considered an extremely unlikely event, genes showing recurrent variants that are present only in multiple affected children but not in any of

their parents can be rapidly selected for further assessment [30]. This study design may be particularly applicable to gene discovery in disorders for which most cases are appear sporadic (that is, the parents are unaffected) but a dominant mode of inheritance is suspected or when substantial locus heterogeneity is expected (as is found in intellectual disability [31], schizophrenia [32] or autism [33], for example). However, it should be noted that, in around 70% of cases where a variant is called in a an affected child that is not present in either parent, this inconsistency is eventually found to be artefactual, related either to a failure to call the variant in one or both parents or, alternatively, to a false-positive sequencing error in the affected child [20].

1.5.3 Effect on Protein Structure as a Means of Variant Filtration

As mentioned in section 1.2, around 85% of all disease-causing mutations identified (prior to the advent of next-generation sequencing) were exonic and thus altered protein sequence. Many of the remaining intronic variants altered canonical splice sites. Therefore, it is common practice to filter out synonymous exonic variants and intronic variants, provided they do not affect canonical splice sites. However, one must always bear in mind that splice site prediction methods are imperfect and it is possible that disease-causing synonymous variants may be inadvertently discarded because they affect a splice site (or, indeed, a transcription start site for an alternative transcript or even a promotor region for a downstream gene) that is not currently recognised as such.

Candidate variants remaining after synonymous variants can then be further stratified on the basis of their predicted deleteriousness to protein structure. In general, greater weight is given to variants that cause a frameshift, introduce a stop codon or disrupt a canonical spice site since such variants tend to result in either a more radical alteration in protein structure than simple missense SNVs or, alternatively, lead to a reduction in quantity or even complete loss of the protein product from the mutated allele, due to nonsense-mediated decay and other such factors. Nonetheless, the assumption underlying this approach should be recognised as an oversimplification. Redundancy in the molecular machinery of the cell means that complete loss of many proteins can be tolerated, even when bialleleic, and may not result in any perceptible impact on health. Indeed, based on a sample of 185 individuals, the 1000 Genomes project

recently reported the average healthy person carries an average of 14.3 to 15.9 biallelic frameshift or stopgain mutations, depending on the population studied [34]. On the other, even a single amino acid substitution might be sufficient to cause disease if it is located in some critical functional domain of the protein or radically alters the proteins final structural conformation.

Occasionally, when decent linkage of homozygosity data are lacking and most of the genome is remains under consideration, investigators may resort to eliminating non-frameshift indels or retaining only truly novel variants in a bid to see if anything 'stands out'. However, since non-frameshift deletions are sometimes pathogenic (i.e. the ΔGAG mutation in *TOR1A*) and variants with reduced penetrance may make their way into databases of normal sequence variation at low frequencies, this strategy entails a significant risk of accidentally eliminating the true causal variant whilst simultaneously increasing the chance of the erroneous assignment of pathogenicity to some other novel, but ultimately harmless, variant. As such, claims of putatively causal variants identified by this manner would be unlikely to be accepted for publication, unless it were supported by evidence of segregation in one or more independent families.

1.5.4 Other Variant Data Used in Supporting Capacity

Of the other data usually assembled automatically on each variant, conservation scores and *in silico* predictions of pathogenicity are often considered the most useful in helping to prioritise variants and to support the argument for any proposed role in the causation of disease. It should be noted, however, that it is not considered good practice – no matter how tempting at times – to formally eliminate variants on the basis of conservation scores or *in silico* predictions of pathogenicity, for the reasons set out below.

Evidence that a genetic site has experienced purifying selection suggests that it is of functional importance to the organism. By extension, measurements of evolutionary conservation can serve as a relatively reliable proxy marker for sites at which an alteration could potentially have deleterious affects on the organism's health. Variants affecting such sites can thus be considered as having a higher *a priori* probability of

being disease-causing, whereas those affecting sites with neutral or higher than expected rates of evolutionary change are much less likely to be so. In fact, approximately 90% of known pathogenic variants affect evolutionarily conserved bases or regions. At the same time, the fact that at least 10% of reported pathogenic mutations must therefore lie in apparently non-conserved regions means these scores can never be used to exclude a variant absolutely. In addition, there is a significant difference in the utility of conservation scores as a guide to pathogenicity when considering dominant versus recessive disease mutations [35]. Bases where disease-causing variants for dominant disease are generally found to be more conserved than those that are affected by variants causing recessive disease. In the latter case, the variant can effectively 'evade' purifying selection while in the heterozygous state.

Evolutionary conservation scores for most sites across the genome are available pre-computed and are typically calculated by one of three programmes – GERP, PhyloP and PhastCons – each of which employs a slightly different algorithm. More information about these programs can be found in sections 2.7.1 – 2.7.3.

Another means of assessing the likelihood that a variant is deleterious involves the use of *in silico* predictions of pathogenicity. The number of published algorithms designed to generate such predictions is constantly expanding, but, at the time that this work was completed, the programmes most commonly used were SIFT/PROVEAN, PolyPhen2 and MutationTaster. Each employs different strategies in its attempt to identify functional sites at which variation is most likely to affect phenotype and these strategies are briefly explained in sections 2.7.4 – 2.7.7. Overall, however, estimates for the sensitivities of such software tools, based on studies using them to distinguish between well-established pathogenic mutations and known polymorphisms, range from approximately 70 to 90%; more worryingly, estimates of their specificities are much wider, ranging from approximately 15% to 80% [36-39]. It is clear, therefore, that these predictions should only ever be considered as 'an educated guessimate' at the true pathogenicity of a variant.

One consequence of the different approaches adopted by all these algorithms is that there exists significant discordance between the predictions generated by different methods. A recent study of five of the most common prediction methods in use (PolyPhen2, SIFT, LTR, MutationTaster, PhyloP) demonstrated that the correlations between scores for identical mutations at the same site from the various prediction methods were mostly weak to moderate [35]. The authors hypothesised that two factors were predominantly responsible for this. Firstly, the set of species used by one method in deriving evolutionary conservation differ considerably from those used by another. Secondly, the set of perfectly conserved sites used for training by one method may also differ from those used by others, due to a disparity in sequence alignments adopted by each method. Thus, in summary, *in silico* predictions of pathogenicty, whilst not without a role in the assessment of the candidate variants, are best regarded as a source of supplementary information alone and, in an ideal world, they should not be used as the basis for variant filtration

Finally, once the number of variants under consideration has been whittled down to at most 20 or less, it is often necessary to curate more detailed information on each of the variants in order both to help eliminate some and highlight others as priority candidates. At this stage the information has to be manually curated and the process can be quite time consuming. Useful data might include expression profiles, protein function, protein structure (particularly location of the variant with regard to predicted functional domains), biological pathway annotations, known gene-disease associations, and previously published papers making mention of the gene. At times, there is an abundance of data on a particular gene; at others, there is none. In truth, only rarely does this second level of data allow complete exclusion of a gene (usually by revealing absence of expression in the tissue of interest). It can, nonetheless, be helpful in prioritising variants for PCR screening in cohorts of other affected individuals based on the strength of the case for each.

1.6 Early Impact and Promise of NGS in Clinical Genetics

The introduction of NGS technologies has ushered in a new age of promise in clinical genetics. Though it remains in its relative infancy and developments continue at a

significant pace, the early impact of these new techniques might nonetheless be said to have constituted somewhat of a revolution for the field. The main areas of impact, as relevant to the work presented herein, are summarised in the sections below.

1.6.1 Identification of Novel Disease Genes

The rate of identification of novel genetic causes for Mendelian phenotypes has accelerated rapidly since the introduction of NGS. A literature search conducted as part of a recent review paper published in 2012 revealed that more than 100 causative genes in various Mendelian disorders have been identified by means of exome sequencing [40] and this number continues to rise. Genetic analysis of autosomal recessive disease is particular well suited to the application of these new techniques, which explains the greater success rates in kindreds exhibiting this mode of inheritance. In all, 61 out of the 108 newly-identified genes exhibited autosomal recessive inheritance (56.5%), 40 of them were transmitted as an autosomal dominant trait (37%), whilst only one X-linked recessive and one X-linked dominant disorder had yielded [40].

Considering the known 'facts' in medical genetics around the time of the NGS technologies came into existence, it is not surprising that whole exome sequencing generated such excitement. As previously stated, around 85% of disease-causing mutations identified at that time were located in protein-coding regions of the genome [8, 9]. Furthermore, most (91.8%) of the variants that alter protein coding are either nonsense/missense (~56%), small insertion/deletions (~24%), splicing (~10%) or regulatory (~1.8%) mutations, all of which are notionally detectable with standard NGS. It followed, therefore, that – in theory at least – whole exome sequencing alone should be able to identify at least 78% of causal variants.

As regards whole genome sequencing, there have, to date, only been a handful of studies that have successfully employed this technique data to identify a novel disease-causing mutation. This is in part due to the relative expense involved in obtaining this increased breadth of coverage. At the same time, and perhaps more importantly, there remain no rigorous solutions to the problems arising from the current dearth of

information about the extent of normal variation in non-protein-coding regions and the relative difficulty in predicting the potential pathogenicity of the observed variants in these regions. However, as the cost of sequencing continues to fall and databases of whole genome data increase in size, it seems likely that whole-genome sequencing will become increasingly common and perhaps even standard in the near future.

1.6.2 Expanding the Phenotype: Genetic Pleiotropy and Disease.

Pleiotropy is a well-defined concept in genetics: it refers to the idea of a single locus that exerts an influence on two or more, sometimes seemingly unrelated, phenotypes [41]. When disease-causing variants occur in such genes only a subsection of the phenotypes may be altered, depending on the position of the mutation within the protein or other genetic or environment factors, leading to variable presentation in the disease state. It has long been recognised that mutations in the same gene can lead to quite different presentations of the same disease or, indeed, even presentations that are normally considered as entirely different diseases. Nonetheless, the increasing use of exome sequencing has led to a growing awareness of the true extent of such phenotypic pleiotropy and evidence is rapidly accumulating that this may be much more common than was previously thought. For instance, *de novo* mutations in *ATP1A3*, a gene normally associated with rapid-onset dystonia parkinsonism (RPD), were recently shown to be the cause of alternating hemiplegia of childhood, a paediatric neurological condition normally considered quite unrelated to RPD [42-44]. Similarly, a homozygous mutation in *ATP13A2*, a gene usually associated with a form of atypical parkinsonism known as Kufor-Rakeb, was recently found to be segregating with disease is a kindred exhibiting neuronal ceroid lipofuscinosis with typical histopathology, suggesting that lysosomal dysfunction may be a common aetiological factor in these two clinically distinct, neurological disorders [45]. Perhaps the most stunning example of such pleiotropy in disease, however, comes from the recent identification by exome sequencing of mutations in *PRRT2* as the cause of paroxysmal kinesogenic dyskinesia (PKD) [46, 47]. It has subsequently been shown that, in addition to typical PKD, mutations in *PRRT2* may also manifest as infantile convulsions with choreoathetosis, benign familial infantile seizures, episodic ataxia, hemiplegic migraine and/or benign paroxysmal torticollis of infancy [48-51]. Moreover, in the case of *PRRT2* mutations, the

resultant phenotype may differ not merely between kindreds carrying different mutations (as might be expected), but also between kindreds carrying the same mutation and even between affected individuals within the same family [52].

1.6.3 Screening for Known Causes of Genetic Disease.

Currently genetic diagnosis in individuals with hereditary disease can be a time-consuming and costly procedure both for the referring clinician and for the laboratory staff at the receiving institution. The first step in the process of genetic diagnosis is to match the patient's clinical phenotype to one of the reported phenotypes associated with a known disease gene for the condition. If the disease phenotype in question is associated with mutations in a single gene alone, the choice of which gene to sequence is straight-forward. However, this is not the case for many disease phenotypes: many, if not most genetic diseases, exhibit considerable locus heterogeneity, meaning that a particular phenotype may be associated with a long list of possible genetic causes, which, after mode of inheritance has been taken into account, can only, at best, be ordered in terms of their relative frequency of occurrence when deciding what to screen. The genetics of Charcot-Marie-Tooth (CMT) disease exemplify this situation perfectly. Over 70 disease genes have been described for CMT and related disorders and, to a large extent, the presentations associated with these various genes are potentially clinically indistinguishable [53]. Testing instead relies on sequential Sanger sequencing of each possible causative gene, which, because of the cost and effort involved, must be done serially rather than in parallel, until either a disease-causing mutation is found or no further possible genes remain to test (taking into account the fact that, for many rarer genes, testing is only available on a research basis).

The application of NGS technology, in an adapted form, to genetic screening promises to significantly speed up this process and, moreover, to reduce costs in the process. Just as the same basic science can be used to sequence either the entire genome or only the protein-coding part thereof, so the protocol for NGS can also be adapted to sequence any user-defined fraction of the genome by the addition of some initial, simple PCR amplification and library preparation steps. So-called 'disease-targeted sequencing' takes advantage of this fact [54]. In this paradigm, primers are designed to amplify all

protein coding regions of all known genes associated with a particular disease phenotype, thus allowing them to be interrogated in parallel (rather than in series) for potential disease causing variants (see section 2.8 for more details). Moreover, with the addition of special indexing primers, samples from different individuals can be multiplexed and sequenced together in a single run. Several NGS platforms, specifically designed for this purpose and complete with integrated on-board bioinformatics, are now on the market, promising low cost, rapid sample-to-sequence turn-around times. These advances hold out the possibility of speedy testing of whole panels of disease genes simultaneously, which may considerably hasten the process of genetic diagnosis while also reducing the workload for staff at the diagnostic laboratory.

Although NGS platforms have a higher sequencing error rate (approximately 1 error per 10,000 bases sequenced at 20x read depth) compared to Sanger sequencing, this is largely offset by the extremely high read depth achieved as a result of the PCR amplification step (usually >100x). Data from our own diagnostics laboratory as well as that of others suggests that that targeted NGS is just as reliable as Sanger sequencing in detecting pathogenic SNVs and small indels. For instance, in a recent study of DNA samples from 84 patients with hereditary cardiomyopathy, all of the 168 variants identified using an NGS-based approach targeted to all 48 known cardiomyopathy genes were subsequently confirmed with Sanger Sequencing [55]. These variants included deletions up to 18 bp and insertions up to 8 bp. No false-negative or false-positive results were obtained for variants selected for confirmation. These results suggest that, at a read depth of at least 30x per nucleotide, there is no bar to the use of disease-targeted NGS in place of Sanger sequencing in a diagnostic setting.

The methodologies employed in targeted NGS also have research applications and these are of direct relevance to the work presented in this thesis. In presenting evidence for a putative novel disease gene, one of the most compelling arguments for its pathogenicity is the demonstration of one or more independent kindreds who carry mutations in the same gene and in which the mutations segregate with the disease in question. To do this, the putative gene must be screened in a large number of genetically undiagnosed individuals exhibiting a similar phenotype. However,

Mendelian diseases are rare to start with and most exhibit considerable genetic heterogeneity, meaning the number of cases due to any particular disease gene can be very low. Thus, the number of individuals that must be screened in order to stand any chance of identifying independent kindreds with segregating mutations in the same gene is often in the hundreds. When the putative gene has just a few exons, the most efficient way of doing this is by Sanger sequencing; however, when a gene is large, Sanger sequencing becomes a costly and extremely time-consuming process. Instead, the same methodology that underlies disease-targeted NGS can be employed to amplify and sequence only the gene in question. Given the ability to multiplex samples in the same experiment, this can dramatically reduce the time taken to sequence a set number of cases in genes with a sizeable number of exons. This technique has direct relevance to the work presented in this thesis and further explanation of its technical aspects can be found in section 2.8.

1.7 Modern Genetic Techniques for Examining Complex Disease

Up to this point, the discussion in this chapter has been largely centred around Mendelian disease and it is certainly the case that most of the work presented herein is directed towards elucidating the genetic cause of disease that is inherited in that fashion. However, as part of my work, I also examined the genetic contribution to complex disease (in this case, Parkinson's disease). It is to the genetic architecture of complex disease and the development of the genetic techniques used to dissect it that I will now turn to in the sections below.

As previously noted, Mendelian inheritance underlies only a tiny fraction of the number of individuals affected by many diseases; in most cases, the disease appears at least to occur sporadically. However, it is common knowledge that some families show a 'predilection' for certain sporadic diseases, albeit without clear Mendelian inheritance. This everyday observation has been borne out by twin studies, which have demonstrated that most conditions – ranging from hypertension and obesity through to Parkinson's disease and multiple sclerosis – show some degree of heritability. This heritability, which might most profitably be thought of as an in-built susceptibility or risk, is conferred by the individual's genetic make-up, which acts in concert with

environmental factors to determine whether he or she will develop a particular phenotype or disease.

Different hypotheses have been proposed to explain how such in-built risk might be encoded in the human genome. Under the 'common disease, common variant' hypothesis, it is postulated that, for common, sporadic disease, multiple variants with a high minor allele frequency (that is, variants that are found commonly in the general population) will be found to modulate the risk of developing the disease. Individually, each variant would impart only a small additional risk but, collectively, they might act additively or synergistically to produce a more significant effect. Testing such a hypothesis does not, theoretically at least, involve a particularly complicated or even novel experimental design: by examining the genetic make-up of a large number of individuals with a particular common disease and comparing it to that of a similarly large number of individuals without the disease, it should be possible to identify those genetic elements mediating risk since they would be expected to occur at a significantly higher frequency in affected cases when statistically compared to unaffected controls. Practically, however, the technology to obtain genome-wide genotyping information on a large number of individuals was initially lacking and association studies of the time were limited to considering a handful of candidate genes at most and utilised only relatively modest sample sizes. From around 2005 onwards, there was a radical change in the state of play: advances in the understanding of human genetic variability ushered in by completion of the Human Genome Project coalesced with concomitant improvements in genotyping technology to permit, for the first time, the rapid, relatively inexpensive and unbiased examination of the whole-genome for disease-associated elements in enormous case-control cohorts. The genome-wide association study (GWAS) was born. In the sections below, the scientific advances that paved the way for this revolution in the genetics of complex disease are briefly detailed.

1.8 SNPs, Chips and Haplotype Blocks

Single nucleotide variants (SNVs) represent genetic differences between individuals that have been introduced into the human population by random mistakes in the DNA replication process over the evolutionary time frame. The error rate for DNA

replication is very low – estimated at around 1.1 to 3 x 10^{-8} mistakes per base per generation. Nonetheless, in view of the vast size of the genome, errors are inevitable and studies suggest that each individual will carry around 70 novel SNVs with respect to their parents [56]. The frequency of such genetic variants in the population is in part determined by their age, as transmission to subsequent generations would be expected to result in a higher and higher frequency over time. However, some variants may be subject to natural selection or random genetic drift, leading to a lower frequency in the human population than might otherwise be anticipated based on their age alone. In general, SNVs with a minor allele frequency of greater than 5% are considered common and are often also referred to as single nucleotide polymorphisms or SNPs for short. SNVs with a minor allele frequency of less than 0.5% are considered rare, whilst those with a minor allele frequency of greater than 0.5% but less than 5% are said to be of intermediate or moderate frequency.

The initial draft of the Human Genome Project in 2000 identified around 1.4 million SNVs. As present, over 50 million SNVs are known and this number is expected to increase further as more genomes are sequenced. These variants represent a pool of genetic variability that could potentially underlie an individual's in-built risk for sporadic disease. However, genotyping every single one of these SNVs was (at least until the event of NGS) impossible in any statistically meaningful number of people and, even with NGS technologies, the cost is still prohibitive. However, it has long been known that the alleles of different SNVs lying in physical proximity on the same chromosome show non-random combinations – that is, the observation of one allele of an SNVs at one position makes it more likely that you will observe a particular allele of a second SNV nearby. This results from the fact that SNVs in close proximity tend not to be separated by chromosomal recombination during meiosis and thus tend to travel together between generations. This phenomenon is called linkage disequilibrium and the blocks of SNVs between recombination hotspots, which tend to travel together from generation to generation, are termed haplotype blocks. From a genomic perspective, linkage disequilibrium operates only over relatively short ranges and haplotype blocks typically do not extend beyond 200 – 300kb.

The international HapMap project, proposed in 2002, set about creating a map of typical human haplotype blocks [57]. Armed with such a map, the number of SNVs that needed to be genotyped to cover the entire genome was drastically reduced. Instead, 'tagging' SNPs could be selected that served as a marker for a particular haplotype block and the genotype of the remaining SNVs constituting the same block derived probabilistically. Subsequent development of such 'imputation' algorithms made it possible to predict the genotype of millions of SNVs based only on the physical genotyping information of a limited set of approximately 500,000 tagging SNPs. At the same time, advances in genetic technology based on immobilised nucleic acid sequences and labelled allele specific oligonucleotide probes made it possible to miniaturise the genotyping process so that it could be performed on the back of a chip and analysed automatically by a computer. The advent of these so-called 'SNP chips' paved the way for rapid, inexpensive and accurate genotyping of large numbers of individuals across their entire genomes and opened the door to the first truly genome-wide association studies (GWAS), which began to appear in the literature with increasing frequency from 2005 onwards.

1.9 Insights from GWAS of Complex Disease

The GWAS approach has provided important information regarding the pathogenesis of complex disease. Firstly, it has delivered evidence to support the hypothesis that at least a subset of the same pathways involved in Mendelian disease is involved in the more common sporadic forms of disease as well. In the case of Parkinson's disease, GWAS have repeatedly demonstrated strong association signals overlying the genes *SNCA* (α-synuclein), *LRRK2* (leucine repeat rich kinase 2), and *MAPT* (tau microtubule associated protein), all three of which are known causes of Mendelian forms of parkinsonism [3, 4, 58]. These findings lend credence to the argument that therapeutics that are developed based on a knowledge of pathways involved in Mendelian disease may prove more widely beneficial to individuals suffering from sporadic forms of the same disease. It this respect, the development of LRRK2 kinase inhibitors represents a promising therapeutic avenue for Parkinson's disease [59].

Secondly, large-scale international GWAS and subsequent metanalyses of their results have demonstrated that numerous genetic loci modulate risk for sporadic disease, even for disorders that were previously considered to have little genetic component, as was the case for Parkinson's disease [60]. Some of these loci overlie genes that might reasonably have been expected to be associated with the particular condition under examination based on the prior understanding of its pathogenesis (examples include the association signal overlying apolipoprotein E in Alzheimer's disease or those overlying genes involved in the immune response in multiple sclerosis [61, 62]). Many association signals, however, overlie genes involved in pathways that might not otherwise have been connected with that condition. Such unexpected associations may provide new clues to the underlying pathogenesis of the disease in question. Examples of such findings include a role for autophagy in Crohn's disease [63], for the complement pathways in age-related macular degeneration [64], and for the central nervous system in obesity [65]. Yet, despite the exciting opportunity provided by the signposting of such novel avenues for exploration, care must be taken to keep in mind that the exact same phenomenon that makes microarray-based GWAS possible – that is, linkage disequilibrium – also means that the SNPs identified as 'top hits' is such studies do not, in all probability, represent the causal SNVs themselves. Instead, it is altogether more likely that these top hits represent markers that are in linkage disequilibrium with the true causal variant. Indeed, the true causal variant may be located at a considerable distance from the implicated risk SNP, especially if that causal variant is itself rare [66]. For this reason, GWAS hits – particularly those where the underlying gene's function cannot immediately be connected to the pathophysiological mechanisms thought to be relevant to that disease – require further exploration using cellular biological techniques. A good example of this kind of follow-through work can be found by considering GWAS findings in Parkinson's disease. An association signal for this condition was identified in multiple GWAS overlying the gene *GAK* (cyclin G associated kinase) [3, 5, 67]. The protein product of this gene acts as a key mediator of endocytic vesicle trafficking by regulating interactions with adaptor proteins and later driving disassembly of the vesicle clathrin coat [68]. Following on from this, it has now been demonstrated experimentally that *GAK* knockdown accentuates protein load and toxicity in cell culture when α-synuclein is overexpressed and decreases cell viability in

rat primary neuron cultures expressing α-synuclein with the Parkinson's disease-associated p.A53T mutation [69]. This new information suggests that targeting endocytic pathways may be a viable approach to combat Parkinson's disease.

Finally, another important finding to arise from GWAS is that the 'common disease, common variant' hypothesis does not explain all of the heritability of a condition. Most GWAS-implicated common SNPs display only modest individual effects on disease risk and, even when considered together, fail to account for all of the estimated genetic risk [70]. For example, although approximately 60 to 80% of AD risk is estimated to derive from genetic factors [71], known genetic risk loci (including the uniquely large effect of apolipoprotein E) account for just half of this variance in genetic susceptibility [72]. In many other conditions, the figure appears to be closer to 20% [73, 74]. It is likely that at least a portion of this missing heritability may be accounted for by an alternative hypothesis, which states that predisposition to common disease may be determined by multiple rare to moderately-rare variants of higher penetrance. The basis for a good example of a genetic risk factor such as this comes not from any particularly exciting or novel advancement in genetic technology, but instead from good, old-fashioned, astute clinical observation. It was noted some time ago by clinicians caring for individuals suffering from Gaucher's disease – an autosomal recessive lysosomal storage disorder caused by biallelic mutations in the gene *GBA* – that individuals with clinically mild disease, as well as their otherwise neurologically-normal parents, appeared to exhibit a higher incidence of Parkinson's disease than would be expected by chance alone [75]. It has now been conclusively demonstrated that rare heterozygous variants in *GBA* act to moderately increase the risk of developing Parkinson's disease.

'Moderately rare, moderately penetrant' variants, such as those found in *GBA*, are poorly represented in GWAS that are based on microarray genotyping of tagging SNPs. Imputation can only provide probabilistic information on genotypes for known low-frequency SNPs that are not directly genotyped and, furthermore, this can only be obtained for populations in which accurate haplotype information is already available. In an attempt to overcome this problem, association studies are already underway that use sequencing data generated by NGS platforms. Such studies represent somewhat of

an intersection between the genetic techniques relevant to Mendelian and complex disease that have already been explored in this introductory chapter. However, whilst the use of NGS to directly genotype all variants allows – theoretically at least – for an assessment of the contribution of rare variants to complex disease, it brings with it its own set of challenges. Particular concerns include the increased sample size required to detect association when so many more variants are under consideration and the added risk of false positives arising from population stratification, which, as mentioned in section 1.3, appears to be the rule where rare variants are concerned. Since the work presented herein is concerned only with association of common polymorphisms, an full exploration of the above-mentioned and other challenges related to rare variant association analyses is beyond the scope of this chapter; interested readers are, however, referred to the recent review article by Lee *et al.*, which covers this subject in depth [76] and also to the final discussion chapter of this thesis, which touches on future directions in genetic research.

1.10 Summary

In this chapter, I have aimed to give a broad overview of the recent scientific advances that have underpinned latest advances in our understanding of the genetic contribution to neurological disease. Thus far, I hope, what has emerged is the idea of a tripartite model of the contribution of genetic variation to disease susceptibility. This model is represented in figure 1 with reference to the common neurodegerative disorder, Parkinson's disease. NGS technologies have been developed and used to identify extremely rare variants of high penetrance that underlie rare Mendelian forms of a disease, whilst DNA microarrays and the statistical techniques inherent to association analyses have been used to examine common variants of low effect-size that underpin some of the hereditability associated with common forms of the same disease. These techniques are directly relevant to most of the work presented herein. More recently, a combination of the two techniques promises to allow the exploration of the contribution of variants of rare to moderate frequency and intermediate penetrance to complex disease, though many challenges remain to overcome.

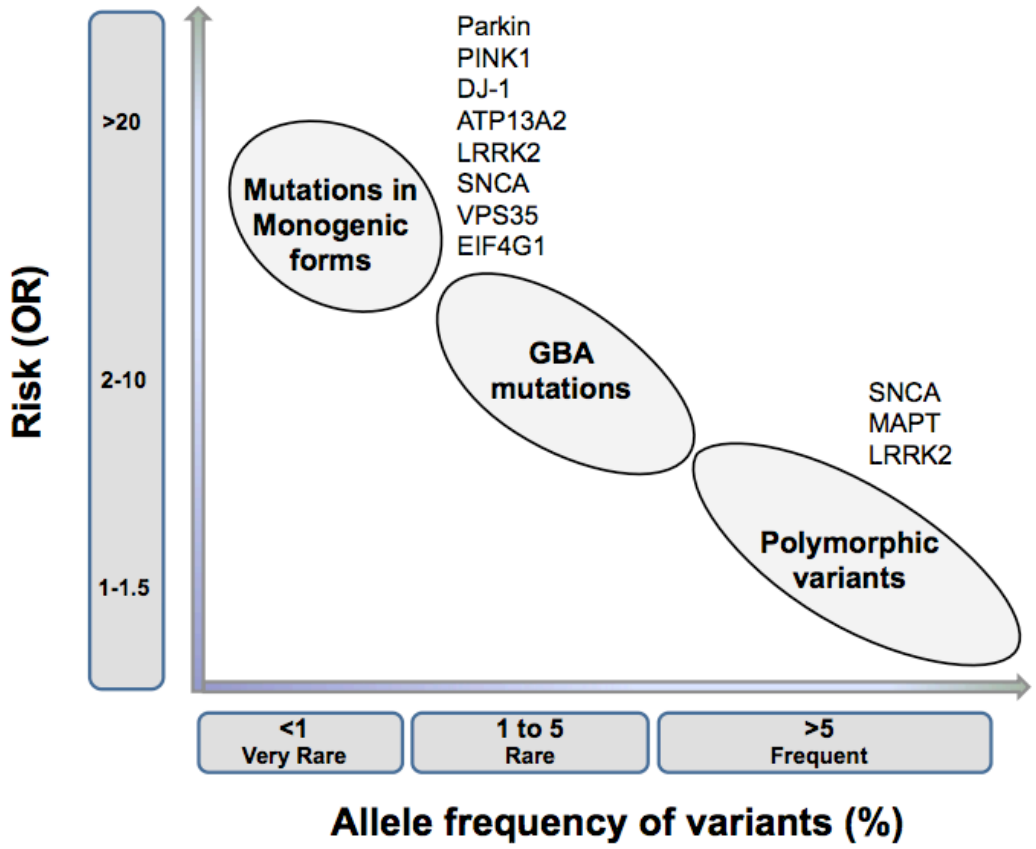

Figure 1 - The tripartite approach to genetic variation as a risk factor for disease, using Parkinson's disease as an example. Very rare Mendelian mutations conferring very high risk are shown on the left; moderately rare mutations of intermediate penetrance are shown in the centre; and some examples of common polymorphisms of small effect are shown on the far right. This figure was originally published in Lesage et al. (2012) [77]

In chapter 2, I first provide some additional technical information that is required to fully understand NGS sequencing technologies and the work exome sequencing work presented herein. This is followed, in chapter 3, by an overview of the current understanding of the genetics of the condition upon which most of my work was focussed – dystonia. Chapter 4 gives details of the core methods used time and again in my research work. Chapters 5 through 15 detail the bulk of my work – exome sequencing in genetically-unresolved kindreds with hereditary neurological disease, a small association analysis of neuropathologically-proven Parkinson's disease and various

PCR-based sequencing experiments aimed at determining the relevance or frequency of proposed Mendelian disease genes. Chapter 16 offers a discussion of the work presented herein and addresses the question of the direction of genetic research is likely to take in the future.

CHAPTER 2:

Technical Aspects of Sequencing and Ancillary Genetic Techniques

2. Technical Aspects of Sequencing and Ancillary Genetic Techniques

2.1 Introduction

In the sections that follow, I discuss some of genetic techniques that I have employed in my attempts to tackle the kindreds detailed in the chapters that follow. The aim here is not to explain the use of these techniques in relation to any particular family (this can be found in the individual chapters that are dedicated to each kindred) but instead to outline the scientific rational that underpins their use and the generic methodology that must be followed to produce accurate and useful data. This repertoire of techniques represent a set of core tools from amongst which the modern geneticist must select the most appropriate combination to elucidate the genetic aetiology of a particular phenotype, based on the characteristics of the disease and the kindred in question.

2.2 Dideoxynucleotide Sequencing

This common, enzymatically-based method of reading small fragments of DNA (≤700bp) was the mainstay of DNA sequencing for around 3 decades since it was first pioneered by Fred Sanger in the mid-70s [7] and is usually termed Sanger Sequencing for short. It relies on the random inhibition of chain elongation by DNA polymerase, creating newly synthesised DNA strands of various lengths that can by separated by size. Random inhibition of chain elongation is achieved by including a small quantity of dideoxynucleotides (ddNTPs), which are similar to their deoxyribonucleotide (dNTP) counterparts but lack the hydroxyl group at the 2' and 3' carbon positions, in the reaction mix. The lack of the hydroxyl group at the 3' carbon in particular precludes the formation of a phosphodiester bond with the subsequent dNTP, preventing DNA polymerase extending the chain any further. Since the input DNA sample is a population of identical molecules population produced by PCR amplification, each of the fragments in one reaction will have a common 5' end (defined by the sequencing primer) but a variable 3' end (defining by the essentially random insertion of the appropriate ddNTP into any one of the many positions that will accept that base). By combining the ddNTPs with fluorescent dyes with different emission wavelengths,

automated DNA sequencing machines could be produced that used laser excitation during electrophoresis to construct the DNA sequence from the intensity profiles of four fluorophores.

The main use of dideoxynucleotide sequencing within this work is to read small portions of DNA – usually no more than a single exon of a gene at a time – in order to confirm the potentially causal variants detected by exome sequencing or, alternatively, to screen the DNA of other individuals with a similar phenotype for mutations in the same exon or gene.

2.3 Next Generation Sequencing

Unlike Sanger sequencing, which is based on the electrophoretic separation of chain-termination products produced in individual sequencing reactions, next-generation sequencing (NGS) involves the massively parallel sequencing of clonally amplified or single DNA molecules that are spatially separated in a flow cell. Sequencing is performed by repeated cycles of polymerase-mediated nucleotide extensions. As a massively parallel process, NGS generates hundreds of gigabases of nucleotide-sequence output in a single instrument run.

Most NGS workflows share the general processing steps delineated in figure 2. As the figure suggests, there is in fact little difference between protocols for whole exome and whole genome sequencing – the former simply includes an additional step involving the ligation of oligonucleotide probes defining the target sequence and the subsequence washing away of non-ligated DNA. Significant differences do exist, however, in terms of the basic chemistry used by different manufacturers of NGS platforms, which has consequences for both library preparation and data generation. In our own lab, we employed Illumina technology using the HiSeq 2000 and details of the chemistry underpinning this technology are given below. For the purposes of comparison only, details of Roche's alternative methodology are also provided, though this was not available within our institution and thus not used to the generate any of data presented in this thesis.

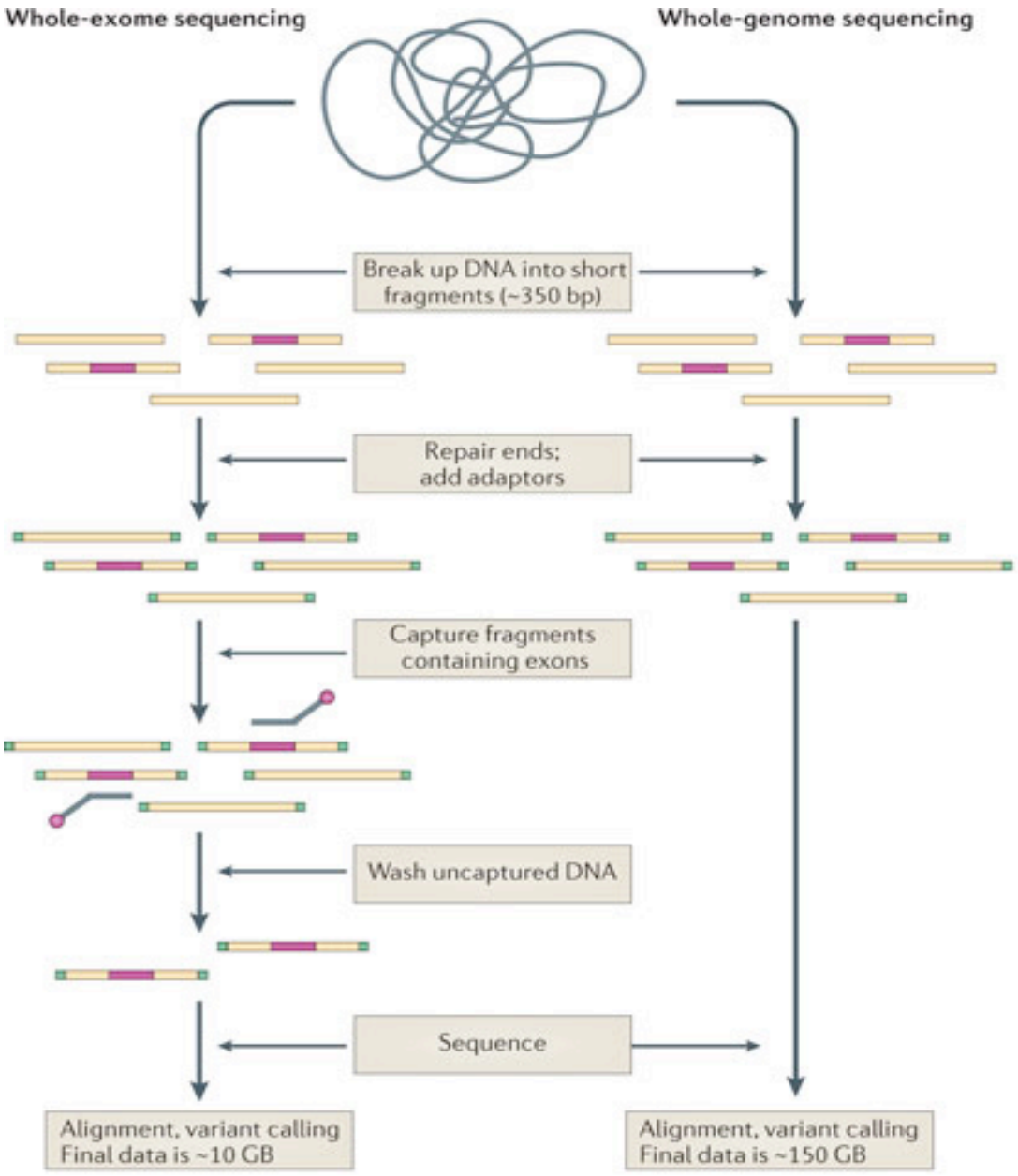

Figure 2 - Next generation sequencing process steps for platforms requiring clonally amplified templates (Roche 454, Illumina and Life Technologies). Input DNA is converted to a sequencing library by fragmentation, end repair, and ligation to platform specific oligonucleotide adapters. Individual library fragments are clonally amplified by either (1) water in oil bead–based emulsion PCR (Roche 454 and Life Technologies) or (2) solid surface bridge amplification (Illumina). Flow cell sequencing of clonal templates generates luminescent or fluorescent images that are algorithmically processed into sequence reads. Figure adapted from Voelkerding et al., (2009) [78]

The first step is to prepare the 'library' comprising DNA fragments ligated to platform specific oligonucleotide adapters. Template DNA is fragmented into lengths of several hundred base pairs and end-repaired to generate 5′-phosphorylated blunt ends. The polymerase activity of Klenow fragment is used to add a single A base to the 3′ end of the blunt phosphorylated DNA fragments. This addition prepares the DNA fragments for ligation to oligonucleotide adapters, which include an overhang of a single T base at their 3′ end to increase ligation efficiency. The adapter oligonucleotides are complementary to anchors attached either directly to the flow-cell (Illumina) or, alternatively, to individual beads (Roche).

The method of clonal amplification differs depending on the system used. On Illumina platforms, the adapter-modified, single-stranded template DNA is added to the flow cell and immobilized by hybridization to the anchors under limiting-dilution conditions. DNA templates are then amplified in the flow cell by 'bridge amplification', which relies on captured DNA strands 'arching over' and hybridizing to an adjacent anchor oligonucleotide. Multiple amplification cycles convert the single-molecule DNA template to a clonally amplified arching cluster, with each cluster containing approximately 1000 clonal molecules. Approximately 50×10^6 separate clusters can be generated per flow cell. On Roche platforms, one the other hand, the library is first diluted to single-molecule concentration, denatured, and then hybridized to individual beads containing sequences complementary to adapter oligonucleotides. The beads are compartmentalized into water-in-oil microvesicles, where clonal expansion of single DNA molecules bound to the beads occurs by means of emulsion PCR. After amplification, the emulsion is disrupted and the beads containing clonally amplified template DNA are enriched. The beads are again separated by limiting dilution and deposited into individual picotiter-plate wells, which function as the flow cell.

Sequencing – the final step in the process – differs again depending on the platform used. On Illumina platforms, the clusters are denatured and a subsequent chemical cleavage reaction and wash leave only forward strands for single-end sequencing.

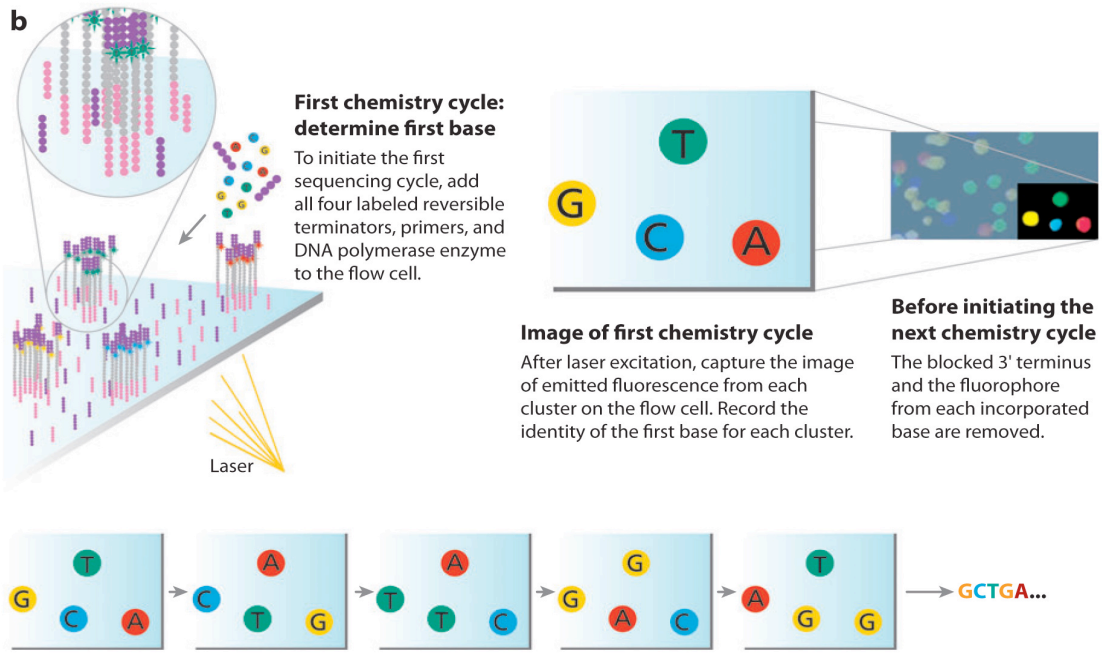

Figure 3 - Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3'-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3'-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation. This figure was published in Mardis et al. (2008) [79]

Sequencing the forward strand is initiated by hybridizing a primer complementary to the adapter sequences, followed by the addition of polymerase and a mixture of four differently coloured fluorescent reversible dye terminators. The terminators are incorporated according to sequence complementarity in each strand in a clonal cluster. After incorporation, excess reagents are washed away, the clusters are optically interrogated, and the fluorescence pattern recorded. With successive chemical steps, the reversible dye terminators are unblocked, the fluorescent labels are cleaved and washed away, and the next sequencing cycle is performed. Roche systems, on the other hand, employ iterative pyrosequencing (successive flow addition of the 4 dNTPs). A

nucleotide-incorporation event in the picotiter plate well containing clonally amplified template produces pyrophosphate release and thus localized luminescence, which is transmitted through the fiber-optic plate and recorded on a charge-coupled device camera. With the flow of each dNTP reagent, all wells are imaged simultaneously, analyzed for their signal-to-noise ratio, filtered according to quality criteria, and subsequently algorithmically translated into a linear sequence output. Regardless of the chemistry used to generate the raw sequence read data, the output is stored in a standardised format as a FASTQ file.

2.4 Post Processing of Exome Data

To generate the final variant calls for analysis, the FASTQ files are fed through an automated bioinformatics pipeline. The data is passed sequentially through a series of freeware tools designed to accomplish the following steps:

1. Verification of input integrity, quality checks, read trimming and primer contamination removal
2. Gapped alignment to the reference genome
3. BAM conversion, sorting and indexing
4. Duplicate removal
5. Local realignment around the position of known indels
6. Quality score recalibration
7. Variant calling
8. Association of variants to annotations recorded in a number of publically available databases (dbSNP, 1000 genomes, Washington exome sequencing project, complete genomics) and to pre-computed *in silico* predictions of pathogenicity and conservation.
9. Data post-processing and output as Excel-compatible, comma-separated files text files for manual inspection and filtration.

Although alternative pipelines exist (utilising different software tools to perform particular tasks such as alignment and variant calling), the actual individual steps involved in the process remain much the same regardless. It should be noted, however, that the selection of one pipeline over another is not without consequences in terms of

the final list of variants that are actually called, as is discussed in section 2.5.3 below. An expanded explanation of the bioinformatics pipeline employed in our own laboratory can be found in section 4.16 .

2.5 Technical Limitations of Exome Sequencing

As with any technology, it is important to be aware of some of the technical limitations of NGS in general and of exome sequencing in particular. Essentially, exome sequencing can be thought of as a modular process, consisting of three core steps that are completed sequentially to produce the final output file. These steps are: 1) library preparation; 2) sequencing; and 3) bioinformatic analysis. For each of these steps, there are important technical considerations that may impact on the likelihood of success for a study aiming to identify a novel disease-causing mutation.

2.5.1 Library Preparation: Defining the Exome

One of the perceived strengths of exome sequencing over the candidate gene studies that it superceded is that there is supposedly no requirement for *a priori* hypotheses regarding the gene(s) likely to be responsible for the disease being tackled. However, this is not quite true: exome sequencing is predicated on the hypothesis that the causal variant will be located within a protein-coding portion of the genome. This is a certainly not an unreasonable hypothesis given that, as mentionned previously, approximately 85% of all causal mutations for Mendelian disease discovered so far have been found in such regions. However, this figure is in all likelihood inflated as a result of detection and publication bias. Firstly, the limitations of scale imposed by the genetic technology available when these disease-causing mutations were identified (i.e. linkage analysis and Sanger sequencing) effectively precluded any detailed examination of non-coding regions due to their size. Secondly, there is a tendency to publish only successful studies, such that the true number of studies in which linkage was obtained but no causal mutation was found will never be known. Thus, it is important to realise that one inherent limitation of exome sequencing is that it runs the risk of missing the causal variant if that variant is not located in a coding sequence and that, furthermore, the proportion of disease-causing variants located in what are traditionally considered non-coding regions may be much higher than anticipated for the reasons detailed

above. Indeed, the recent identification by two independent groups of a hexanucleotide repeat expansion in the 5' region of the gene *C9orf72* as the cause of hereditary amyotrophic lateral sclerosis (ALS) and fronto-temporal lobar dementia (FTLD) serves to illustrates this point well [80, 81]. Not only is this expansion located in a non-coding portion of the genome and thus would not be captured by exome sequencing, but the expansion range is typically in the 100s to 1000s and would thus be too large for detection by Sanger sequencing. Nonetheless, this mutation appears to be the single most common genetic cause of this spectrum of neurodegenerative disorders that has been identified to date, with pathological expansions having been observed at frequencies of up to 29%, 50%, and 88% in case series of FTLD, ALS, and FTLD-ALS, respectively [82].

The situation is further complicated by the fact that the exome is, in practice at least, a somewhat arbitrary and changeable concept. Theoretically, the exome is defined as the entire protein-coding portion of the human genome. However, in spite of the rapid advances in our genetic understanding, there remains considerable uncertainty about which genomic sequences are actually protein-coding and which are not. From an operational perspective, the exome is often defined by the content of the consensus coding sequence (CCDS) database, a consortium effort to delineate the human sequences for which there is strong evidence of transcription and translation. This database currently contains 25,564 unique IDs, corresponding to 18,407 human genes. Even if this were the only database in existence, the exome would still remain necessarily fluid as knowledge of which sequences within the genome are transcribed and translated is constantly improving, leading to changes in the content of the CCDS database. However, the waters are further muddied by the existence of other alternative databases of coding sequences. Some, such as RefSeq, include sequences with a lower level of evidence for coding and are correspondingly larger, whilst others, such as GENCODE, are archival and thus contain numerous redundant sequence identifiers. It is likely that these non-consensus databases contain many more errors than the relatively conservative CCDS database, but it is equally likely that they contain a large number of sequences not yet included in the CCDS database that will, in fact, turn out to be protein coding.

This uncertainty regarding the exome is reflected in the variability of targeting between the various commercially available exome capture kits. By way of example, table 2 (taken from a recent paper by Coonrod *et al.* [2012] [83]), shows comparative metrics for some of the most commonly used kits currently on the market.

Table 2 - Comparative metrics for three commonly used exome capture kits used in library preparation for exome sequencing by 2012. CCDS, RefSeq and GENCODE databases are described in the text above. RefSeq plus = RefSeq exons + 5' and 3' UTRs + microRNA + noncoding RNA. miRBase = database of microRNAs sequences. Rfam = database of RNA families.

Capture Kit	NimbleGen SeqCap v2.0	Agilent SureSelect 50Mb Kit	Illumina TruSeq
Target Region Size	44.1Mb	50Mb	62Mb
Number of Targets	~300,000	331,518	201,121
Number of Probes	2.1 million	430,400	340,427
Probe Type	DNA	RNA	DNA
Bases covered at 10x	81%	>85%	85%
On target reads	90% (±100bp)	77% (±200bp)	~83% (±150bp)
Databases for probe design	CCDS RefSeq miRBase	CCDS RefSeq miRBase GENCODE Rfam	CCDS RefSeq RefSeq plus GENCODE Predicted microRNAs

On the whole, over the three-year period that this research spanned, we observed a general tendency for the size of the targeted region to expand, as more databases were included in the probe design process, and for the depth of coverage obtained to progressively improve. The exome capture kits used at the start of this research project targeted only ~30Mb and coverage was often suboptimum; by the end of the project, the Illumina TruSeq kit, which targets a genomic area twice that size and generally produces excellent coverage, had become the standard choice. Clearly, kits with larger target areas stand a higher chance of including sequences where a causal variant could potentially lie and thus a study's chance of success may be determined by the target

used for library preparation. To date, at least one study has been reported in which the disease-causative variant was initially missed because it was located in a sequence not defined as protein coding by the CCDS; it was only later identified when the larger RefSeq definition of the exome was used as the basis for capture [84].

2.5.2 Sequencing Failures: Breadth and Depth of Coverage

It is worth noting two important technical limitations of exome sequence that are evident from the table above, namely: 1) only a portion of the target will sequence successfully (breadth of coverage); and 2) that the number of reads covering any particular successfully-sequenced target is uneven and can be suboptimal (depth of coverage). Each NGS platform has its own particular characteristics related to the underlying chemistry employed in sequencing that determines which regions are more likely to be poorly represented or more prone to sequencing errors. In general, however, all have a tendency to perform poorly in regions of DNA with certain physiochemical properties, such as those with high GC content or those consisting of highly repetitive or low complexity sequences of nucleic acids. Sequences such as these present difficulties in probe design, probe binding, PCR amplification and subsequent mapping of the resultant sequence to reference genome and can thus result in little or no coverage of areas that could nonetheless potentially habour a disease-causing mutation.

Low coverage depth is a particular problem when the causal variant is likely to be heterozygous. If only a few reads are obtained for a sequence that contains a heterozygous disease-causing variant, there is a significantly increased probability that all the reads obtained may come from the strand containing the normal allele and thus the variant would be missed even though the area had apparently been covered. This problem compounded by the fact that all exome sequencing kits exhibit reference bias, as capture probes are designed to match the reference sequence and thus tend to preferentially enrich the reference allele at heterozygous sites [85]. If the variant is located on the anti-sense strand to the reference, it is much less likely to be detected at low read depths. Moreover, NGS technologies rely on redundancy of coverage to compensate for a relatively high intrinsic error rate, which approaches 1% for most platforms.

When considered alone, a sequencing error is indistinguishable from a sequence variant. However, even if all reads contain a 1% variant-error rate, the combination of eight identical reads that cover the location of the variant will produce a strongly supported variant call with an associated error rate of 10^{-16} [86, 87]. Thus, when evaluating an exome sequencing study, it is necessary to consider not just the breadth of coverage, but also the depth of coverage. At present, there is a general consensus that good quality exome data will cover around 80% of target region at a minimum depth of 10 reads. Studies with coverage metrics that fall significantly below these values have a far greater chance of false-negative and false-positive variant calls and any conclusions drawn from such data should be evaluated with caution.

2.5.3 Limitations of Sequencing Data Analysis

Processing of raw exome sequencing data is a computational intensive process and the algorithms involved in these steps (delineated in sections 2.4 and 4.16) are subject to constant change and development. The first step in this process involves alignment of the sequencing reads to the reference genome. The choice of alignment algorithm will impact on the final coverage values, as different algorithms show varying false-positive and false-negative rates [88, 89]. Even the best mapping algorithms will fail to align all reads to the reference genome. This may be due to sequencing errors, structural rearrangements or insertions in the query genome, or deletions in the reference. Furthermore, it is not possible to unambiguously assign reads to all genomic regions. The 'mappability' of a sequence within a genome has a major influence on the average final mapped read depth and is an important source of false-negative and false-positive single-nucleotide variant calls [90]. Since sequence reads that map to multiple sites in the genome are usually discarded, genomic regions with high sequence degeneracy or repetition show a lower mapped read coverage – inversely correlated to the size of the genomic repeats – than do unique regions, creating a form of systematic bias [91]. In addition, pseudogenes, which show high sequence homology to their expressed counterparts but are subject to greater genetic drift and thus more likely to contain variants, also pose a problem for read mapping. As the surrounding sequence can be essentially identical, a sequencing read containing a variant that actually originated from the pseudogene may be incorrectly mapped to its expressed counterpart,

introducing false-positive errors. Such false positive SNVs were observed in a recent study using exome sequencing in patients with Fanconi anaemia, a condition for which some of the known causative genes have closely-homologous pseudogenes, forcing the authors to resort to long-range PCR to prove the true their true origin [92].

To some extent the problem of low mappability arises as a result of the short read lengths used by many NGS platforms, including the Illumina platform used in our own laboratory. Longer read lengths increase the chance of a read encompassing a unique sequence that anchors all remaining sequences. However, for read lengths greater than 200bp, the gains in mappability fall off rapidly and a recent study has demonstrated that even a read length of 1000bp would still not permit the unique alignment of reads for many coding regions of human genes [93]. A second approach to this problem has been to generate paired-end libraries with longer insert sizes, which increases the chance of one read of the pair mapping to a unique region outside the repeat sequence. This approach is employed by Illumina in its HiSeq 2000 platform, which utilises paired end reads of approximately 100bp in length.

Regardless of the chemistry used in performing NGS, there still remains considerable discrepancy between the output of difference data analysis pipelines, even when used with near-default parameterization on the same raw sequencing data. This was beautifully illustrated by a recent study, which passed the same raw sequencing data from 15 exomes generated on the Illumina HiSeq 2000, through five separate variant analysis pipelines (SOAP, BWA-GATK, BWA-SNVer, GNUMAP, and BWA-SAMtools) [94]. In this study, SNV concordance between the five pipelines across all 15 exomes was found to be only 57.4%, whilst between 0.5 to 5.1% of variants were called as unique to each pipeline. More worryingly from a disease gene discovery point of view was the finding that for, novel SNVs (i.e. those not found in dbSNP135), the overall concordance rate was just 11.4%. Indel calling showed even higher levels of discrepancy, with the overall concordance for known and novel indels being just 43.3% and 4.7%, respectively. In a second phase of the research, the authors focused on two of the most commonly used pipelines, based around GATK and SOAP, in order to assess the validity of the variants detected. A total of 1,140 SNVs found in a single

exome were selected for MiSeq validation; 760 of these SNVs were randomly selected from the set of SNVs that were unique to the GATK and SOAP pipelines (380 from each pipeline). The MiSeq's reliance on PCR amplification (as opposed to exon capture), its longer read lengths, and the much higher depth of coverage made its use a strong method of validation for SNVs and indels alike. Of the 1,140 SNVs targeted for MiSeq validation, 919 (81.0%) were successfully amplified and sequenced, with an average read depth of 5,392. Validation rates for unique-to-GATK SNVs were high, with 306 of 315 (97.1%,) being successfully validated. For unique-to-SOAP SNVs, 174 of 289 SNVs (60.2%) were validated. Validation rates for indels were considerably lower: only 180 of 336 (54.0%) unique-to-GATK and 148 of 332 (44.6%) unique-to-SOAP indels could be validated. In this study, the authors conclude that this discordance between pipelines is likely to be the result of many factors including differences in alignment methods, post-alignment data processing, parameterization efficacy of alignment and variant-calling algorithms, and the underlying models utilized by the variant-calling algorithm(s). Indeed, it is clear from this data that each variant-calling pipeline detects variants that others do not and, whilst the accuracy of these *dis*cordant variants is naturally reduced, between 60.2 – 97.1% of unique SNVs and 44.6 – 54% of indels were nonetheless subsequently validated. However, in the realm of biomedical research, missing even a single variant can mean the difference between discovering a disease-causing mutation or not [95].

2.6 From Potential Causal Variants to Disease-Causing Mutation: Confirmation, Segregation and Independent Kindreds

As a result of the inherent limitations of NGS sequencing and its supporting bioinformatics, each variant is best viewed as a hypothesis to be tested. As such, any suspected causal variants must first be confirmed by Sanger sequencing to ensure that they are not artefactual. This is particularly important with indel calls, which are more prone to error, but is, in reality, a requirement for all variants before any further sequencing or experimental work can be undertaken. Subsequently, segregation of the variant with the disease in the index family must be demonstrated by Sanger sequencing in all affected and unaffected individuals for whom DNA is available. At this stage, it is necessary to keep the possibility mind that segregation may not be

perfect as a result of potential issues such as incomplete penetrance and, occasionally, phenocopies. Finally, if segregation is demonstrated in the index family, one would normally then go on to sequence first the exon and, subsequently, if feasible, the entire protein-coding portion of the gene in which the variant lies in a large cohort of individuals with phenotypically similar disease. Typically, the number of individuals that need to be sequenced to stand a reasonable chance of finding further potentially disease-causing variants in the same exon/gene for rare Mendelian diseases will be in the hundreds. Should further individuals who carry the same or a different potentially-causative variant be identified in this cohort, an attempt is made to contact these individuals and establish the pattern of disease in their own families and, provided consent is obtained, to collect DNA samples from other affected and unaffected members of their families in order to try and demonstrate segregation in one or more independent kindreds. Only rarely is experimental work undertaken without having demonstrated segregation in at least one other independent kindred.

2.7 A Basic Guide to Common Variant Annotations

In the sections below, I give a brief overview of some of the annotations of evolutionary conservation and predicted pathogenicity to which frequent reference is made in this work. The programs that produce these annotations often rely on different algorithms, which each take into account a slightly different set of factors in producing their final 'score'. My aim here is in no way to provide a comprehensive account of the internal workings of these algorithms (which is beyond my own understanding), but simply to allow the reader to familiarise themselves with their purpose and the meaning of their individual score systems.

2.7.1 Evolutionary Conservation

Three main programmes are used to assess the evolutionary conservation of a variant: GERP [96, 97], PhyloP [98] and PhastCons [99]. In general, evolutionary conservation scores appear to be one of the most reliable markers of potential pathogenicity of a variant. Approximately 90% of known pathogenic variants affect evolutionarily conserved bases or regions. However, the fact that at least 10% must therefore lie in apparently

unconserved regions means these scores can never be used to exclude a variant absolutely.

2.7.2 GERP (Genomic Evolutionary Rate Profiling)

For each aligned site, GERP defines a 'rejected substitution' (RS) score by estimating the actual number of substitutions at that site and subtracting it from the number expected assuming neutrality (~5.82 substitutions per site). Selectively constrained sites tolerate fewer substitutions than neutral sites and have positive RS scores. This concept was first described in Cooper *et al.* (2005) [96]. GERP++, as described in Davydov *et al.* (2010) [97], is a development of the same methodology, which uses a more rigorous set of algorithms to calculate site-specific RS scores and to thus detect evolutionarily constrained elements.

Each site is scored independently. Positive scores represent a substitution deficit (i.e. fewer substitutions than the average neutral site) and thus indicate that a site may be under evolutionary constraint. Negative scores indicate that a site is probably evolving neutrally; negative scores should not be interpreted as evidence of accelerated rates of evolution because of too many strong confounders, such as alignment uncertainty or rate variance. Positive scores scale with the level of constraint, such that the greater the score, the greater the level of evolutionary constraint inferred to be acting on that site. Pre-computed scores for 35 mammals are available on the UCSC browsers, ranging from a maximum of 6.18 down to -12.36. In general, variants with a GERP score of >2 are felt to have a much greater chance of being pathogenic.

2.7.3 PhastCons and PhyloP

PhastCons utilises a hidden Markov model-based method that estimates the probability that each nucleotide belongs to a conserved element, based on the multiple alignment. PhastCons values may thus vary between 0 and 1 (the closer the value is to 1, the more probable the nucleotide is conserved). It considers not just each individual alignment column, but also its flanking columns.

By contrast, PhyloP separately measures conservation at individual columns, ignoring

the effects of their neighbours. Its score range is similar to GERP (-14 to 6), with positive scores indicating conservation.

The two methods have different strengths and weaknesses. PhastCons is sensitive to 'runs' of conserved sites, and is therefore effective for picking out conserved elements. PhyloP, on the other hand, is more appropriate for evaluating signatures of selection at particular nucleotides.

2.7.4 In Silico Predictions of Pathogenicity

To date, several variant prioritization tools have been developed that aim to identify damaging alleles. The most commonly used and most robustly tested are SIFT (Sorting Intolerant From Tolerant) and PolyPhen (Polymorphism Phenotyping). Experiments using these tools to try to correctly predict known pathogenic and benign variants have demonstrated that the sensitivity of SIFT and PolyPhen is reasonably high (~70% in both cases), but that their specificity may be as low as 13% and 16%, respectively [36, 38, 39]. Both algorithms proved to be significantly better at predicting loss-of-function mutations than gain-of-function mutations. For the purpose of this thesis, I have used these tried-and-tested favourites, alongside newer programs, such as MutationTaster, which aims to improve upon their basic approach by integrating more information into the calculation, and PROVEAN, which is similar to SIFT but can additionally make predictions for indels.

2.7.5 SIFT and PROVEAN

SIFT [100] employs multiple sequence alignments to assay conservation levels of novel amino acid changing variants with the underlying assumption that sequence variants that alter highly conserved positions in protein sequences are *a priori* more likely to be damaging. SIFT scores range from 1 to 0, with scores less than 0.05 corresponding to a prediction that the substitution will be damaging. The authors suggest the tool has a sensitivity of 85.03% and a specificity of 68.95%, leading to an overall accuracy of 74.77%, though, as mentioned above, others studies using this algorithm to predict known pathogenic or benign variants suggest that these figures – particularly the specificity – are overinflated.

PROVEAN (Protein Variant Effect Analyser) [101] uses a similar approach to SIFT, based on sequence conservation, but can also handle small indels. The procedure used to generate a score in PROVEAN is complicated and for the purposes of this work, it suffices only to know that if the score is equal to or below a predefined threshold of -2.5, the protein variant is predicted to have a 'deleterious' effect. If the PROVEAN score is above this threshold, then the variant is predicted to have a 'neutral' effect. The authors suggest that PROVEAN has a slightly lower sensitivity of 79.76%, but an improved specificity of 78.63%, leading to an overall accuracy of 79.19%.

2.7.6 PolyPhen-2

PolyPhen-2 [102] uses eight sequence-based and three structure-based predictive features that were selected automatically by an iterative greedy algorithm. The majority of these features involve comparison of a property of the wild-type (ancestral, normal) allele and the corresponding property of the mutant (derived, disease-causing) allele, which together define an amino acid replacement. The most informative features characterize how well the two human alleles fit into the pattern of amino acid replacements within the multiple sequence alignment of homologous proteins, how distant the protein harbouring the deviation is from the wide-type protein, and whether the mutant allele originated at a hypermutable site. The alignment pipeline selects the set of homologous sequences for the analysis using a clustering algorithm and then constructs and refines their multiple alignment. The functional significance of an allele replacement is predicted from its individual features by a Bayes classifier.

PolyPhen-2 assigns a qualitative label to the variation – either 'benign', 'possibly damaging' and 'probably damaging' – and calculates the Bayesian posterior probability that this mutation is damaging. Thus, scores closer to 1 indicate a greater likelihood that the variation is damaging. The authors suggest the tool has a sensitivity of 88.68% and a specificity of 62.45%, leading to an overall accuracy of 75.56%. Again, as for SIFT, other studies suggest that these values, particularly the specificity, may be overestimated.

2.7.7 MutationTaster

MutationTaster [103] also employs a Bayes classifier to predict the potential pathogenicity of a genetic alteration, but its strength lies in the breadth of the inputs that the classifier receives. Examples of such inputs include evolutionary conservation scores (PhyloP and PhastCons), interspecies nucleotide and amino acid conservation, splice site predictions (NNSplice), annotated protein domain disruption, and likelihoods of an alteration resulting in protein truncation or non-sense mediated decay, amongst others.

The Bayes classifier is fed with the outcome of all tests and the features of the alterations and calculates probabilities for the alteration to be either 'disease-causing' or a harmless 'polymorphism'. The prediction is accompanied by a 'probability value', which indicates the security of the prediction, i.e. a value close to 1 indicates a high 'security' of the prediction. The authors suggest the tool has a sensitivity of 87.9% and a specificity of 89.3%, leading to an overall accuracy of 88.6%.

2.8 Targeted Next-Generation Sequencing

Targeted high-throughput sequencing brings the power of NGS techniques to bear on the problem of sequencing small areas of interest within the genome to a high degree of accuracy with a rapid turnaround time. The platform used in our own laboratory for this purpose – and the current market leader in this field – is the Illumina MiSeq and this machine thus forms the basis of the discussion below.

As with Sanger sequencing, the individual areas for sequencing (known as amplicons) in targeted NGS are small, typically around the size of an average human exon (150 - 250bp). However, up to 1,536 amplicons can by sequenced simultaneously from a single DNA sample and, by multiplexing samples using special indexing primers, up to 198 DNA samples can be included in a single experiment, resulting in an enormous saving in the time and hands-on lab work required to generate the same amount of data when compared with traditional Sanger sequencing. The turnaround time – regardless of the number of amplicons or samples included – is an invariable 3 days from DNA to data. For the purpose of comparison, taking into account constraints of equipment

availability within a typical laboratory such as our own, a researcher using Sanger sequencing might be able to produce, on average, data for two exons in 198 samples per day. The total time taken to sequence even a single 30-exon gene in 198 individuals would thus be 15 working days (probably closer to 20 when primer optimisation and inevitable repeats due to primer or reaction failures are taken into account). Even this simple scenario demonstrates the huge savings in time that the new technology allows. Moreover, these savings only increase the greater the number of amplicons that are included per MiSeq experiment. For example, it would take an experimenter, using traditional methodologies and working at the same pace as above, two years (including weekends) to produce the same amount of data as could be obtained from a single 3-day MiSeq experiment when running at full capacity.

Moreover, this enormous saving in labour and time is achieved without any significant reduction in accuracy. As mentioned in section 1.5.3, studies comparing the accuracy of targeted high-throughput sequencing with traditional deoxynuxleotide sequencing for the purpose of diagnostic genetic testing have shown no significant difference in their rates of mutation detection or generation of false positives, leading to the conclusion that targeted high-throughput sequencing of panels of disease genes will soon become the standard methodology used in diagnostic genetic laboratories.

As regards cost, targeted high-throughput sequencing is slightly cheaper than deoxynucleotide sequencing, but still represents a considerable outlay, especially in terms of research work. At the time of writing, an experiment generating data for 30 amplicons in 198 individuals currently costs around £3,500. Within the context of this thesis, the main use of this new technique has thus been to screen large numbers of DNA samples quickly for mutations in sizable genes of interest, where traditional sequencing would be extremely time consuming and costly. In addition, in our own laboratory at least, it was felt that at least one other confirmatory mutation in an independent pedigree should have been observed before NGS sequencing of the remaining portions of the gene might be considered a financially sound investment.

In terms of the actual underlying chemistry, the Miseq platform relies on the same techniques of cluster generation, bridge amplification and optic interegation of incorporated flurophores in order to generate the sequencing data, but the initial library preparation requires a few additional steps (see figure 4 overleaf). Firstly, small areas of interest (amplicons) are targeted by hybridisation of custom oligonucleotide probes to unfragmented genomic DNA. After the removal of unbound oligonucleotides, the intervening sequence between probes is synthesised by a DNA polymerase, which begins extension from the upstream oligonucleotide probe; subsequently a DNA ligase contects the extended sequence to the downstream oligonucleotide probe, resulting in the formation of products containing the area of the targeted regions of interest, flanked by sequences required for amplification. Next, PCR amplification is used to drive the incorporation of indexed sequencing primers (which identify the individual DNA samples using a simple grid referencing technique and also include common adaptors required for cluster generation). These first steps take about a day to complete.

Ampure XP beads are then used to separate the PCR products from other reaction components and the samples are normalised to ensure an equal sample representation in the final library. Once purified and normalised, the earlier inclusion of indexing sequences means the samples can be pooled into a single tube without risk of sample cross-contamination. After a brief heat denaturation step, the pooled library is ready for loading into the MiSeq flow cell. This second part of the process also takes about a day to complete.

Finally, generation of clusters by bridge amplification, sequential incorporation of flurophores, optic interrogation and data generation is completed automatically on the MiSeq over a period of around 18 hours, meaning that the data is ready for subsequent analysis by the end of the third working day.

As the sample preparation for the MiSeq includes a PCR amplification step, coverage in number of reads per base tends to be much higher (often > 100x) than for whole exome sequencing (~20 – 50x), without any lost of accuracy.

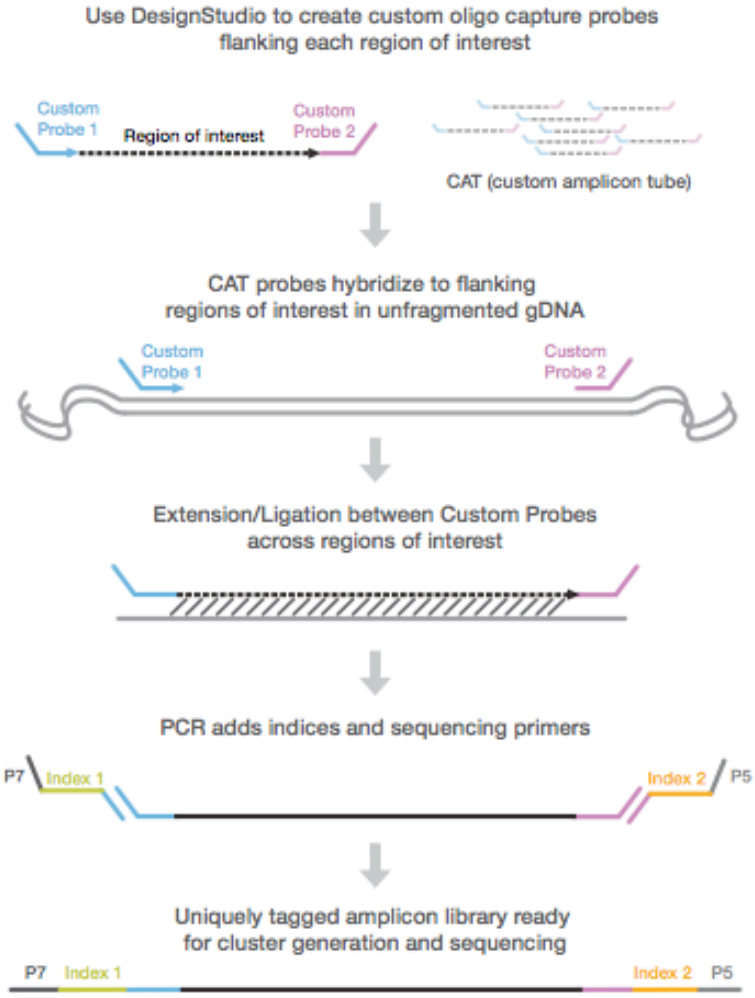

Figure 4 – A schematic representation of the initial library preparation steps in targeted NGS using Illumina's MiSeq. The target sequences are first targeted by hybridisation of custom oligonucleotide probes created with aid of Illumina's DesignStudio software. Extension by DNA polymerase adds the sequences required for amplification. During PCR amplification, special indexing primers are incorporated onto the ends of the products that permit the multiplexing of samples during later stages of the process without risk of sample cross-contamination.

2.9 Linkage Analysis

Until around 5 years ago, genetic linkage analysis combined with positional cloning was the primary tool used in trying to establish the causal genetic variant underlying familial diseases. Even today, in the age of NGS, it remains of paramount importance in

helping to narrow down the number of potential variants under consideration by 'sign-posting' areas of the genome where the causal variant is likely to lie.

In essence, genetic linkage analysis is a statistical method that is used to associate functionality or phenotype of a gene or variant with their location on chromosomes. It relies on the phenomenon of recombination during meiosis and knowledge of the position of genetic markers spread throughout the genome. Since recombination is a monotone, but non-linear function of the distance between two any two genetic loci, the position of a disease gene can be inferred by finding markers that tend not to recombine in all affected individuals.

The probability that an odd-number of crossovers will occur between two loci during meiosis (and that the loci will thus be physically separated by recombination) is termed termed the recombination fraction and represented by the Greek letter θ. This value is defined by the Haldane map function and ranges from 0 (complete linkage) to 0.5 (no linkage whatsoever). On average, about one recombination event occurs every 10^8 bases and this distance is defined as 1 Morgan. For closely linked loci (where $\theta < 0.1$), the recombination fraction approximates the genetic map distance, such that two loci showing recombination in 1% of meiosis are approximately 1cM apart. These basic concepts underpin the complex mathematics of linkage analysis.

In practice, the result of linkage analysis is a table of LOD scores that can be used to identify one (or several) candidate regions, which may still contain many potentially causative genes. The LOD score is derived by a maximum likelihood analysis, which calculates the probability that two loci are linked, expressed as a logarithmic function of the odds ratio favouring linkage [104]. Conventionally, a LOD score of >3, which indicates a probability of 1000 to 1 that the loci are linked, is taken to be statistically sound evidence of linkage. Conversely, a LOD score of -2, which indicates a probability of 100 to 1 that the loci are not linked, is considered sufficient to rule out linkage between two loci. The mathematics of linkage analysis is computationally intensive, especially when pedigree is large or when information is incomplete and inferences must be made regarding missing genotypes and phase. Several pieces of software exist

to perform the analysis with one of the most common being MERLIN (Multipoint Engine for Rapid Likelihood Inference), described in Abecasis et al., 2002 [105].

2.10 Autozygosity Mapping

Mapping of autosomal recessive disorders is more problematical than for autosomal dominant or X linked disorders. Many autosomal recessive disorders are individually rare, making it difficult to collect a sufficient number of cases unless this is done on an international collaborative basis. In addition, in most parts of the world family sizes are limited, with it being uncommon for families to have more than three to four children and therefore making it unusual for there to be more than one or two affected individuals within a sibship.

In 1953, Smith observed that offspring of consanguineous matings would be homozygous for genetic markers near the disease gene [106]. Over 30 years later, Lander and Botstein subsequently realised that a recessive trait could thus be mapped using the offspring of such consanguineous matings, an approach they termed 'homozygosity mapping', but which is probably more correctly termed autozygosity mapping [107]. The principle behind autozygosity mapping is that a fraction of the genome of any particular offspring of a consanguineous mating is expected to be homozygous by virtue of identity by descent. On average, one-sixteenth of the genome of offspring of first-cousin matings would be expected to be homozygous. Furthermore, one might anticipate that the regions of homozygosity should be randomly distributed throughout genome in the different offspring of these matings, except at a common disease locus shared by all affected offspring. Thus, with the use of the genotyping data from the offspring of several first-cousin matings, markers can be identified that are linked to a recessive disorder by virtue of the fact that they are homozygous for the same allele in all affected individuals.

This method works best when several small families from the same population are used together to define the regions of interest, provided there is not significant genetic heterogeneity in the cause of the disease under investigation. However, the approach

can also be used successfully in single families, especially if combined with exome sequencing data.

2.11 Summary

In this chapter, I have tried to provide an overview of the technical aspects of the genetic techniques commonly applied in attempting to eluciadate the cause of genetic disease in a kindred. In addition, I have highlighted some of the limitations of these techniques, which both the reader and researcher must be aware of when interpreting data generated by these means. When considered collectively with the information provided in chapter 1, there emerges, I hope, the concept of a general workflow – incorporating numerous pieces of supporting information – that the researcher follows in order to try and move from the raw exome data to the disease causing mutation. This workflow, which underpins all of my work on exome sequencing, is summarised graphically in figure 5 overleaf.

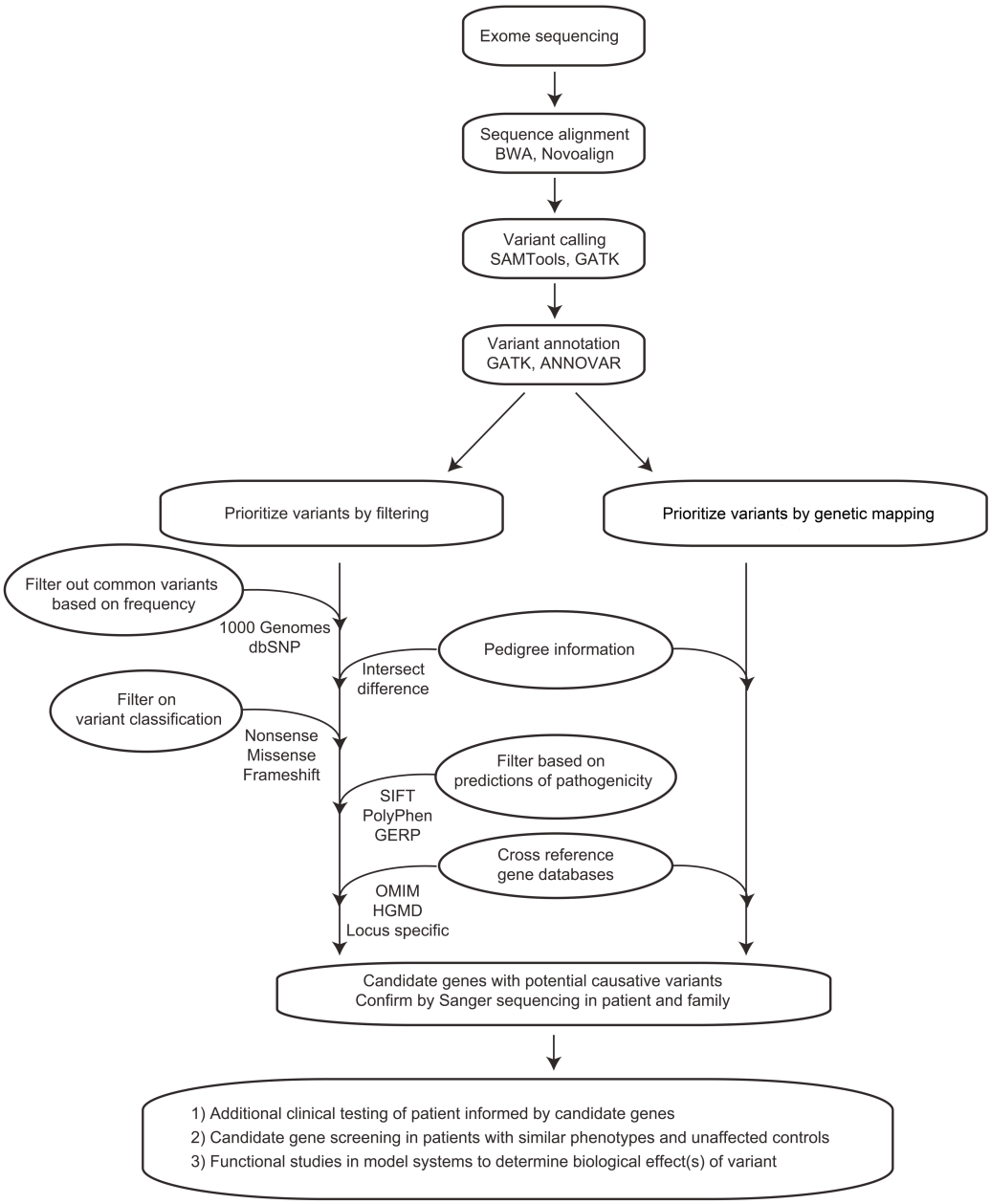

Figure 5 - A schematic representation of the typical workflow that is followed when trying to identify a disease-causing mutation in a kindred with hereditary neurological disease of unknown cause by use of exome sequencing and genetic mapping techniques (linkage analysis/autozygosity mapping). The process involves the drawing together of numerous strands of information used to whittle down the list of potentially causal variants. In addition, the process does not stop once a candidate disease-causing mutation has been identified. Further sequencing in the index family and phenotypically similar case cohorts as well as, where possible, functional studies to assess the biological effect of the variant are prerequisites if the findings are to merit publication. This figure was adapted from Sims *et al.*, 2014 [86].

CHAPTER 3:

Clinical, Genetic and Molecular Aspects of Dystonia

3. Clinical, Genetic and Molecular Aspects of Dystonia†

3.1 Introduction

In the sections of my thesis dealing with the use of exome sequencing, the majority of the patients that I worked with had a form of hereditary dystonia. In this chapter, therefore, I present an overview of the state of knowledge with regard to the genetics and molecular pathophysiology of dystonia as it stood during the time of my PhD. It should be noted that a number of the genes discussed in this chapter were only identified as a cause of hereditary dystonia during the active period of my research and this discussion does not therefore represent a thus not represent an accurate picture of the understanding in this field when I originally took up my post. Moreover, discussion of the genes that I identified within my PhD is not included here, but can be found in chapters 5 and 7. They are, nonetheless, included in tables and figures.

The dystonias are a heterogenous group of hyperkinetic movement disorders, characterized by involuntary sustained muscle contractions affecting one or more sites of the body that lead to twisting and repetitive movements or abnormal postures of the affected body part. It is the third most common movement disorder worldwide [109-112]. Approximately 70,000 people are affected by dystonia in the UK alone, including some 8,000 children and adolescents [113]. Affected individuals can suffer considerable physical and psychosocial distress, which has been demonstrated to have a significant impact on their quality of life [114-116]. The pathophysiology of the disorder is poorly understood at present. In all probability, it is heterogeneous and arises from a dysfunction of the various central neural circuits that control and coordinate voluntary movements, such as those found in the basal ganglia, the cerebellum, the sensorimotor cortex, and the interactions between these three regions of the brain [117-120].

3.2 Classification of Dystonia

The classification of the dystonia is complex and not entirely satisfactory. Several approaches or systems operate in parallel. Clinically, the dystonias are usually classified

† The work presented in this chapter was peer-reviewed and published with adaptations in Charlesworth *et al.*, 2013 [108].

according to one of 4 major variables: 1) age of onset (early onset vs. adult onset); 2) distribution of affected body parts (focal, multifocal, segmental or generalised); 3) the underlying cause (primary, secondary or heredodegenerative); or 4) special clinical features (paroxysmal, exercise-induced, task-specific or dopa-responsive) [121]. The current European Federation of Neurological Societies classification scheme is a two-axis approach that incorporates many of these factors (see table 3) [121].

From a genetic point of view, hereditary dystonia can be classified either by the gene causing the condition, where it is known, or by reference to one of the ever expanding list of dystonia loci, of which there are currently 23 (see table 4). The system of DYT loci is particularly unsatisfactory. The system was designed to indicate genomic regions that had been linked to a specific hereditary disorder, but where the actual causative gene was not yet known [122]. DYT loci were assigned in chronological order based on the appearance of reports in the medical literature. In theory, once the underlying genetic cause was known, the locus was supposed to be withdrawn and the disorder merged into the entry for the cloned gene. However, in practice, this has not happened and both clinicians and researchers alike tend to use dystonia loci and genes names interchangeably, e.g. DYT1/*TOR1A* or DYT6/*THAP1*. With time, several other problems have arisen. The designation of some loci has never been replicated and is of questionable significance (e.g. DYT7 or DYT13), whilst others are known to be the result of incorrect assignations due to erroneous linkage (e.g. DYT9, DYT14 and DYT19) [22, 23]. Some DYT loci did not even designate any chromosomal location, but were (prior to this work, at least) based solely on the observation of a few families with a similar phenotype or mode of inheritance (e.g. DYT2) [123, 124]. More importantly, not all pathogenic mutations causing dystonia have been assigned to a DYT locus (e.g mutations in *SPR*, *CIZ1* or *GNAL*), whilst some syndromes with prominent dystonic components have been assigned to loci belonging to other movement disorders (e.g. PARK13 and PARK14) or *vice versa* (DYT3 and DYT12).

Table 3 - Classification of dystonia, based on the European Federation of Neurological Societies current 2013 scheme.

Axis I – Clinical Characteristics	
Age at Onset	
• Infancy	Birth to 2 years
• Childhood	3 years to 12 years
• Adolescence	13 to 20 years
• Early adulthood	21 to 40 years
• Late adulthood	> 40 years
Distribution of Affected Body Parts	
• Focal	Single body region (e.g. writer's cramp, blepharospasm)
• Segmental	Contiguous body regions (e.g. cranial and cervical, cervical and upper limb)
• Multifocal	Noncontiguous body regions (e.g. upper and lower limb, cranial and upper limb)
• Generalised	Both legs and at least one other body region (usually one or both arms)
• Hemidystonia	Half of the body
Temporal Pattern	
• Disease Course	Static vs progressive
• Variability	Persistent, action-specific, diurnal or paroxysmal
Associated Features	
• Evidence of another movement disorder	Isolated if dystonia +/- tremor only (previously termed 'pure'); or combined if dystonia and another movement disorder present
• Evidence of other neurological or systemic manifestations	List of co-occurring neurological manifestations
Axis II – Etiology	
Nervous System Pathology	
• Evidence of Degeneration	Previously termed heredodegenerative dystonia
• Evidence of Structural Lesion	Stroke, tumour, demyelination ect
• No Evidence of Above	Previously and often still termed 'primary' dystonia
Inherited or Acquired	
• Inherited	Autosomal dominant/recessive, X-linked, or mitochondrial
• Acquired	Perinatal brain injury, infection, drug-induced, toxic, vascular, neoplastic, brain injury or psychogenic
• Idiopathic	Sporadic or familial

Table 4 - The current DYT loci with brief description of associated phenotype, gene of linkage interval (where known), mode of inheritance and OMIM reference numbers. Note: DYT9, DYT14, DYT19 are not included in the table as they are now known to be synonymous with DYT18, DYT5a, and DYT19 respectively. Inaddition, there is great confusion over what the symbol DYT23 actually refers to and it has thus been omitted.

Locus Symbol	Phenotype	Gene or linkage (if known)	Mode of Inheritence	OMIM
DYT1	Early-onset primary torsion dystonia	*TOR1A*	AD	605204
DYT2	Early-onset primary dystonia with prominent cranio-cervical involvement	See chapter 7	AR	224500
DYT3	Adult onset dystonia-parkinsonism, prevalent in the Philippines.	*TAF1*	X-Linked	31420
DYT4	Whispering dystonia (adult onset spasmodic dysphonia) with generalisation and 'hobby horse' gait	*TUBB4A*	AD	128101
DYT5a	Progressive dopa-responsive dystonia with diurnal variation	*GCH1*	AD	128230

DYT5b	Akinetic rigid syndrome with dopa-responsive dystonia or complex encephalopathy	*TH*	AR	191290
DYT6	Adult-onset torsion dystonia with prominent cranio-cervical and laryngeal involvement	*THAP1*	AD	602629
DYT7	Adult-onset primary cervical dystonia	18p	AD	602124
DYT8	Paroxysmal non-kinesigenic dyskinesia	*MR1*	AD	118800
DYT10	Paroxysmal kinesigenic dyskinesia	*PRRT2*	AD	128200
DYT11	Myoclonic dystonia (often with alcohol responsiveness)	*SGCE*	AD	159900
DYT12	Rapid onset dystonia parkinsonism and alternating hemiplegia of childhood	*ATP1A3*	AD (often de novo)	128235
DYT13	Early onset torsion dystonia in one Italian family	1p36.32-p36.13	AD	607671
DYT15	Myoclonic dystonia with alcohol responsiveness in one Canadian kindred	18p11	AD	607488
DYT16	Early-onset dystonia-parkinsonism	*PRKRA*	AR	612067

DYT17	Primary focal dystonia with progression in one Lebanese family	20p11.2-q13.12	AR	612406
DYT18	Paroxysmal exercise-induced dyskinesia +/- epilepsy	*SLC2A1*	AD	612126
DYT20	Paroxysmal nonkinesiogenic dyskinesia 2, in one large Canadian family	2q31	AD	611147
DYT21	Adult-onset mixed dystonia with generalisation in one Swedish family	2q14.3-q21.3	AD	614588
DYT22	Reserved, but not published	?	?	?
DYT24	Tremulous cranio-cervical dystonia +/- upper limb tremor without generalisation	See chapter 5	AD	610110
DYT25	Craniocervical predominant primary dystonia with generalisation +/- hyposmia	*GNAL*	AD	615073

3.3 Investigation of Dystonia

The diagnosis of dystonia is, fundamentally, clinical. It relies on the presence of repetitive or sustained abnormal postures (with or without tremor) and the recognition of specific features, such as a *geste antagoniste* or overflow and mirror movements. *Geste antagonist* refers to a voluntary manoeuvre (such as touching the face or an affected body part) that temporarily reduces the severity of dystonic posture or movements. An overflow movement is an unintentional muscle contraction which accompanies, but is anatomically distinct, from the primary dystonic movement, i.e. posturing of a hand normally unaffected by dystonia when performing tasks with the affected hand. Conversely, mirror movements are dystonic postures of a body part normally affected by dystonia when performing a motor task with a body part that is not affected by dystonia.

In general, for primary dystonia, few, if any, tests are required. The main exception to this rule is early-onset (<30 years of age) dystonia of unknown aetiology, which should always prompt consideration of a diagnosis of dopa-responisve dystonia or Wilson's disease, as accurate identification of these diseases at an early stage will permit the introduction potentially life-changing treatments. Therefore, many would advocate, at the very least, a metabolic analysis that includes measurement of serum copper and caeruloplasmin and, possibly, a trial of L-dopa in this group. In practice, an MRI and genetic testing for *TOR1A* and *THAP1* mutations are often also performed. Finally, given the recent identification of a new form of treatable dystonia caused by brain manganese deposition secondary to mutations in *SLC30A10* [125, 126], serum manganese measurement should at least be considered.

The features listed in table 5 are those that might raise suspicion that the dystonia is not primary and trigger further investigation. The purpose of such investigations is to identify a secondary cause for the dystonia or to further elucidate the cause of dystonia presenting as part of a heredodegenerative condition. In practice, a combination of blood tests, structural imaging and selected secondary investigations are usually required to secure the diagnosis and table 6 gives an indication of some of the investigations that may be appropriate given the aetiology under consideration. As

regards neuroimaging, MRI is generally the modality of choice, although secondary CT may be required to accurately distinguish calcium from iron deposition in the basal ganglia. A dopamine transporter (DaT) scan may be useful to distinguish dopa-responsive dystonia or rapid-onet dystonia-parkinsonism (where it will be normal) from other causes of parkinsonism with secondary dystonia [127, 128].

Table 5 - Some features that should raise suspicion that dystonia is secondary or heredodegenerative and trigger further investigation. It should be noted that some of these features are found in some types of primary dystonia, but their presence should nonetheless trigger careful consideration of a secondary dystonia or heredodegenerative disorder.

Features suggestive of non-primary dystonia
Abnormal birth or perinatal history
Dysmorphia
Delayed developmental milestones
Seizures
Hemidystonia
Sudden onset or rapidly progressive dystonia
Prominent oro-bulbar dystonia
The presence of another movement disorder (except tremor)
Neurological signs suggesting involvement of other neurological systems (pyramidal signs, cerebellar signs, neuropathy, cognitive decline)
Signs suggesting disease outside of the nervous system (hepatomegaly, splenomegaly)

Table 6 - Some investigations used in the diagnosis of secondary and heredodegenerative dystonia with example indication. Please note: this list is not exhaustive.

Investigation	Example indications
Blood	
Acanthocytes	Neuroacanthocytosis, neuronal brain iron accumulation
Alpha fetoprotein	Ataxia telangiectasia
Creatinine Kinase	Neuroacanthocytosis
Copper and caeruloplasmin	Wilson's disease, neuronal brain iron accumulation
Lactate and pyruvate	Mitochondrial disorders
Serum manganese	Dystonia with brain manganese deposition due to *SLC30A10* mutation
Serum ferritin	Neuroferritinopathy
White cell enzymes	Lysosomal storage disorders
Urine	
Urinary aminoacids	Aminoacidaemias
24h urinary copper	Wilson's disease
Neuroimaging	
MRI	Most secondary causes, looking for structural lesions, iron/calcium deposition, caudate atropy, white matter abnormalities, etc.
Dopamine Transporter (DaT) Scan	Parkinsonism
Other	
Nerve condition tests	Spinocerebellar ataxia, neuroacanthocytosis, metachromatic leukodystrophy
Trial of L-dopa	Early onset dystonia (<30 years of age) of unknown aetiology
Autonomic function tests	Multiple system atrophy
Slit-lamp examination	Wilson's disease
Liver biopsy	Wilson's disease
Muscle biopsy	Mitochondrial disorders
Electro-retinography	Neuronal brain iron accumulation
Genetic tests	See table 9 for the genetic causes of heredodegenerative dystonias

3.4 Genetic Burden in Dystonia and Genetic Testing

Current evidence suggests that there is a significant genetic contribution to many forms of dystonia. Monogenic inheritance is most often seen in early-onset cases, where a family history can often be elicited. However, the reduced penetrance of some monogenic forms of dystonia, such as those due to mutations in *TOR1A* and *THAP1*, means that many apparently sporadic cases may also fall into this category. Furthermore, it is likely that a number of genes responsible for familial dystonia remain to be discovered, such that negative genetic testing for all currently known dystonia genes does not imply that the disorder is not genetic.

Late-onset dystonia, which represents by far the greatest number of cases, also appears to have a strong genetic basis. Studies based on the clinical examination of first-degree relatives of patients with focal dystonia have reported a risk of developing the same or another form of dystonia in the range of 23 to 36% [129, 130]. Epidemiological studies have suggested that, although often apparently sporadic, adult onset dystonia may sometimes be inherited in an autosomal dominant manner, but with a markedly penetrance of 12 to 15% [129, 131, 132]. Unfortunately, this presents significant challenges for gene discovery and, at present, the genetic architecture of late onset dystonia remains largely unknown. It is possible that the use of endophenotypes, based on imaging or neurophysiological testing, to identify non-penetrant mutation carriers will help by permitting accurate linkage and segregation analyses in some of the larger kindreds [110, 133]. However, multigenic inheritance, resulting from the combination of two or more genetic changes, each imparting a low to moderately increased risk of developing dystonia and acting in combination with environmental factors, may also underlie a significant proportion of the apparent heritability of late-onset dystonia. Large-scale genome wide association studies will be helpful in dissecting out the genetic contribution in these cases.

At the present time, genetic testing is most profitably employed in familial or early-onset cases, where it is likely to have the highest yield. Despite the fact that the establishment of a molecular diagnosis rarely alters management radically, it can be helpful for patients to understand the cause of their dystonia and also bring a halt to

unnecessary continued investigation, as well as allowing the physician to impart accurate information regarding the risk of recurrence in subsequent generations. A scheme for deciding which genes may be appropriate to test in which patients is provided below in figure 6

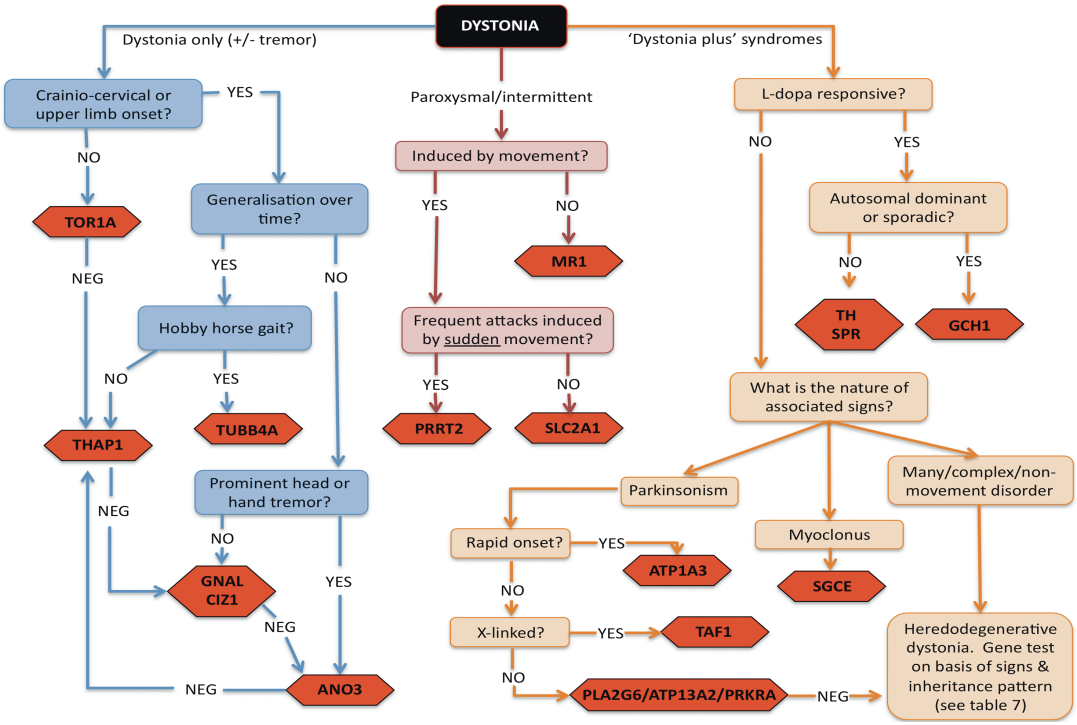

Figure 6 - A workable strategy for identifying the likely genetic basis for some major forms of dystonia. Mutational screening for some genes is currently only available on a research basis

3.5 Monogenic Forms of Dystonia

In the sections below, we present an overview of the major forms of dystonia exhibiting Mendelian inheritance for which an underlying genetic cause has been identified. Table 7 offers an overview of some of the key features of each of the various subtypes of primary dystonia where the causative gene is known.

Table 7 - Summary of the key clinical features of some of the major forms of primary dystonia by causative gene. Features described relate to typical cases and it must be borne in mind that there is often wide phenotypic variation between cases. † = clinical description restricted to rapid-onset dystonia-parkinsonism only (for alternating hemiplegia of childhood, see main text)

	Gene	Typical age range for onset	Other notable non-dystonic signs	Typical distribution of dystonia (at onset)	Generalisation or progression?	Clinical clues or other special features
ISOLATED DYSTONIAS	*TOR1A* (DYT1)	Childhood	-	Legs >> arms	Often Generalises	Jewish ancestry; dystonia progressing to fixed deformity; laryngeal or cranial sparing
	DYT2 See chapter 7	Childhood to adolescence		Craniocervical	Yes	Autosomal recessive inheritance pattern or consanguinity
	THAP1 (DYT6)	Adolescence to early adulthood	-	Laryngeal, cervical or brachial	Generalisation common	Laryngeal or brachial onset with progression.
	CIZ1	Adult	-	Cervical	No generalisation	Only reported in pure focal cervical dystonia
	DYT24 See chapter 5	Adolescence to early adulthood	-	Craniocervical or brachial	No generalisation	Prominent head, voice or arm tremor.
	TUBB4A (DYT4)	Adolescence to early adulthood	-	Laryngeal, craniocervical	Generalisation observed	Ataxic, hobbyhorse gait; extrusional tongue dystonia; single family
	GNAL (DYT25)	Adolescence to mid-life	-	Craniocervical	Generalisation observed	-

PARAOXYSMAL DYSTONIAS	*MR1* (DYT8)	Childhood to adolescence	Choreatic dyskinesias; ballism	Limbs and face	No	Long (mins to hours) attacks, triggered by alcohol, caffeine and emotional stress.
	SLC2A1 (DYT18)	Childhood to adolescence	Epilepsy (see clinical clues also)	Legs	Progression to static dystonia can occur	Exercise or hunger induced. May be associated with other complex neurology – ataxia, spasticity ect – or haemolytic anaemia.
	PRRT2 (DYT10)	Childhood to adolescence	Choreatic dyskinesias; childhood epilepsy.	Limbs (generally the limb affected is the limb that is moving)	No	Frequent, brief (secs to mins) attacks triggered by movement. Family history of childhood epilepsy or hemiplegic migraine
COMBINED DYSTONIAS	GC*H1* (DYT5a)	Childhood	Parkinsonism in some	Legs	Progressive, but often becomes static later.	Diurnal variation; dramatic response to L-dopa; may mimic a spastic paraparesis.
	TH (DYT5b)	Infancy to childhood	Progressive hypokinetic rigidity; complex encephalopathy in some; 'lethargy-irritability' crises.	Generalised but often with axial hypotonia.	Progressive without treatment	L-dopa response related to phenotype: response tends to be less good in those with complex encephalopathy.
	SPR	Infancy to childhood	Motor/speech delay, parkinsonism, dysarthria, autonomic symptoms, oculogyric crises	Generalised but often with axial hypotonia.	Progressive without treatment	L-dopa response related to phenotype: response tends to be less good in those with complex encephalopathy. 5-HTP may be a useful adjunct.
	SGCE (DYT11)	Childhood	Myoclonus. Neuropsychiatric manifestations	Cervical	Dystonia usually remains segmental	Myoclonus usually more prominent than dystonia. Often alcohol responsive
	ATP1A3† (DYT12)	Adolescence to adulthood	Parkinsonism	Rostocaudal gradient	Rapid onset then stabilises	Onset in context of febrile illness or other stressor. No response to L-dopa.
	PRKRA (DYT16)	Childhood	Variable parkinsonism	Any body part	Generalisation	No response to L-dopa

3.6 The Primary Isolated Dystonias

Primary isolated dystonia is usually defined as a syndrome in which dystonia is the only clinical feature (except for tremor of the arms or head and neck) without any evidence of neurodegeneration or any obvious secondary cause (e.g. trauma, autoimmune, post-infectious, ect.).

3.6.1 TOR1A Mutations (DYT1/Oppenheim's Disease)

In 1911, Oppenheim proposed the term 'dystonia musculorum deformans' to describe a syndrome in children with twisting or jerking movements, muscular spasm, gait abnormalities that often progressed to a fixed postural deformity. The hereditary nature of the condition was subsequently recognised and, in 1990, the disease was linked to chromosome 9q32-34 and designated with the first dystonia locus (DYT1) [122]. It was quickly realised that this locus was associated with a significant proportion of all childhood-onset dystonia, particularly in Ashkenazi Jews [134, 135]. Seven years later the gene responsible for the condition was identified and named *TOR1A* [136]. The *TOR1A* gene comprises 5 exons and is widely expressed. In almost all cases of DTY1-related dystonia, the underlying genetic mutation is an inframe GAG deletion located in exon 5 of the gene, resulting in the loss of a single glutamic acid in the final protein product (TorsinA). This mutation has been shown to be responsible for around 80% of all primary, early-onset dystonia in Ashkenazi Jewish populations and up to 50% of primary, early-onset dystonia in non-Jewish populations [137, 138]. Only two other missense variants (p.A288G and p.F205I) and one frameshift, 4-base-pair deletion (c.934_937delAGAG) in this gene have since been reported to cause dystonia, though their pathogenicity is by no means certain [139-141].

The penetrance of the GAG deletion is notably low, with current estimates somewhere within the region of 30% - 40% [134, 137]. Moreover, even when the condition does manifest, the spectrum of severity of the symptoms is wide, ranging from mild focal dystonia to severe and disabling generalised dystonia. The exact mechanism underlying this variability remains unclear, but is presumed that other genetic or environmental modifiers must exist that influence the penetrance and presentation of the disease. To date, only one coding polymorphism in exon 4 of *TOR1A* (rs1801968), which encodes

either aspartic acid (D) or histidine (H), has convincingly been shown to modify the risk of developing symptoms of dystonia. It appears that carrying the H allele *in trans* (that is, on the opposite allele from that harbouring the GAG deletion) may reduce the risk of developing symptoms of dystonia by ten-fold, to as little as 3% [142, 143]. Interestingly, in cellular models, the overexpression of *TOR1A* containing the H allele *in cis* was sufficient by itself to induce formation of inclusions of torsinA in a manner similar to expression of GAG-deleted *TOR1A*, though at a much lower level [144]. It is thought that this may be the result of a disruption of protein interactions induced by the presence of the exposed histidine residue. However, expression of GAG-deleted *TOR1A* containing the H allele actually reduced the formation of inclusions, implying that these two changes act to cancel each other out to some degree [144]. It has even been suggested that the GAG-deletion may need to be carried in conjunction with a D allele *in cis* to be penetrant [143], though this has not yet been proven. Yet, although the effect of the 216H coding polymorphism appears to be robust (*in trans*, at least), its relatively low population frequency (approximately 12% in Europeans and less in other populations) means it cannot fully explain the reduced penetrance or variable expressivity seen in this condition; other factors clearly remain to be identified.

In terms of nongenetic factors that might affect the penetrance of TOR1A mutations, one small study has assessed exposure to perinatal adversity, childhood infections, general anaesthesia and trauma in manifesting and non-manifesting carriers as well as non-carriers from a series of 28 families with DYT1-related dystonia by means of a retrospective questionnaire [145]. Only self-reported perinatal adversity, in particular complications of vaginal delivery, showed a statistically significant ($p=0.02$) positive correlation with manifestation of the disease.

Clinically, DYT1-related dystonia typically presents in childhood with dystonic posturing of the foot or leg, though it may begin in any part of the body, and evolves to generalised dystonia with fixed deformities. As mentioned above, however, late-onset or much milder forms of the disorder are recognised [138]. There is generally a family history consistent with autosomal dominant inheritance, albeit with reduced penetrance, but apparently sporadic early and late onset cases are also seen [146].

The *TOR1A* gene encodes a protein TorsinA, which is a member of the AAA+ family of ATPases and is believed to be involved in maintenance of both structural integrity and/or normal function of protein processing and trafficking. TorsinA is expressed throughout the central nervous system in humans, but is found at particularly high levels in the dopaminergic neurones of substantia nigra pars compacta, locus coeruleus, Purkinje cells, cerebellar dentate nucleus, basis pontis, thalamus, hippocampal formation, oculomotor nuclei and frontal cortex [147-149]. The exact mechanisms by which torsinA causes dystonia are still unclear, but there is accumulating data to suggest that torsinA is important for cellular compartments and pathways, including the cytoskeleton, the nuclear envelope, the secretory pathway and the synaptic vesicle machinery [150-155].

3.6.2 THAP1 Mutations (DYT6)

After the discovery of *TOR1A* and the subsequent uptake of genetic testing for mutations in this gene, it became clear that a second, distinct form of pure primary dystonia existed. Affected individuals, who were negative for *TOR1A* mutations, tended to have an older age at onset - in their adolescence or adulthood - and presented predominantly with cranio-cervical or focal dystonia affecting the upper limbs. In 1997, the DYT6 locus was assigned to chromosome 8p21-q22 by linkage analysis in two Mennonite families, but it was not until 2009 that the gene responsible for the condition was finally identified as *THAP1* (thanatos-associated protein domain-containing apoptosis-associated protein 1) [156, 157]. Missense, nonsense and frameshift mutations, spread throughout most of the coding portion of the gene, have all been associated with disease. Mutations in this gene have been detected in genetically diverse populations throughout the world and, unlike *TOR1A*, do not seem to be especially prevalent in any one particular population. Inheritance is autosomal dominant with a reduced penetrance. To date, penetrance has only been measured in Amish-Mennonite families, where it appears to be around 60%, but this may not be true of all populations or mutations [158]. Most mutations are found in the heterozygous state, but homozygotes are described [157, 159].

The clinical spectrum of *THAP1* mutations is wide. Presentation with oromandibular, cranio-cervical or laryngeal dystonia is common, but presentations with focal dystonia of the limbs, segmental or generalised dystonia are all described in the literature. In a recent analysis by Xiromerisiou *et al.* of 100 patients reported to carry *THAP1* mutations, the mean age at onset of dystonia was 24 years, 60% of patients were females and the distribution was generalised in 37%, segmental in 30%, multifocal in 6% and focal in 27% [160].

The THAP1 protein is an atypical zinc finger protein characterized by an N-terminal THAP domain, a proline rich region and a c-terminal nuclear localization domain [161]. The THAP domain is conserved in both vertebrates and invertebrates [162]. It has DNA binding properties and is thought to be involved in the regulation of transcription, either on its own or in concert with other proteins. Thus, one possible mechanism by which mutations in *THAP1* might cause dystonia is by dysregulated transcription of key genes. THAP1 protein is also known to interact with the prostate apoptosis response protein 4 (PAR-4), which is a transcription regulator involved in apoptosis [161]. Under conditions of cellular stress or toxicity, this protein is rapidly upregulated and it has been linked to neurodegeneration in a number of disease models [163-165]. One intriguing finding has been that THAP1 may regulate transcription of *TOR1A* by binding to one of two sites in its promotor region [166, 167]. Mutations in *THAP1* have been shown to disrupt binding to the *TOR1A* promotor and decrease *TOR1A*-driven luciferase expression, thus suggesting that under some conditions *TOR1A* is negatively regulated by THAP1 protein and that mutations in *THAP1* may lead to abnormally high levels of torsinA [168]. Although these data come from experimental cell lines and thus are of uncertain physiological significance, this does at least raise the possibility that the disease mechanisms in DYT1 and DYT6 dystonia may be linked. However, if this is the case, the question arises as to how to square this mechanistically with the prevailing idea that the GAG-deletion in *TOR1A* leads to a loss of function of the protein [153]. One possibility is that *TOR1A* expression must be maintained within a set range in relation to its binding partners and that dysfunction may result from either increased expression or loss of function [169]. Indeed, transgenic mice overexpressing wildtype

TOR1A have been shown to display evidence of neurohistological, neurochemical and behavioural abnormalities, which may provide some support for this hypothesis [170].

3.6.3 GNAL Mutations (DYT25)

During the period of this research, mutations in the gene, *GNAL*, were identified as further cause of familial primary cervical dystonia [171]. Using whole exome sequencing and linkage in two unrelated families, two mutations in GNAL were initially identified, a nonsense mutation (p.Ser293*) resulting in a premature stop codon in one family and a missense mutation (p.Val137Met) in the other. Screening of 39 additional families identified 6 further novel mutations, with segregation confirmed in all families for which DNA from additional affected individuals was available.

Clinically, of the 28 individuals with dystonia resulting from a mutation in GNAL, 82% had onset in the region of the neck and 93% had cervical involvement when examined for the study. However, progression to other sites had occurred in at least half of those affected and generalised dystonia was seen in about 10% of cases. Thus, the clinical phenotype in *GNAL*-related dystonia appears to be not dissimilar to that caused by mutations in *THAP1*. Though the authors noted that onset in *GNAL* mutations was never brachial, this may just be a chance finding given the small number of cases known so far, rather than a true distinguishing factor.

GNAL encodes $G\alpha_{olf}$, the alpha subunit of triheteromeric G protein G_{olf}, which is involved in dopamine (D1) signaling. Experimental evidence exists to show that $G\alpha_{olf}$ is mostly responsible for the coupling of D1 receptors to adenylyl cyclase in striatal neurons and that $G\alpha_{olf}$ is required for D1-mediated behavior and biochemical effects [172-174]. Since D1 dopamine receptors have a known role in mediating locomotor activity, the link between *GNAL* and dystonia is biologically plausible. The gene is located on the short arm of chromosome 18 in a region that has been linked not only to DYT7 dystonia (though the original DYT7 family do not appear to carry a mutation in *GNAL*), but also to bipolar disorder, schizophrenia and attention deficit hyperactivity disorder [175-179]. Interestingly, homozygous knockout mice for this gene are hyposmic and display hyperactive behaviours [180].

3.6.4 CIZ1 Mutations

In early 2012, Xiao et al. reported that they had used a combination of linkage analysis and whole exome sequencing to identify a mutation in *CIZ1* (p.Ser264Gly) as the likely causal variant in a large Caucasian kindred with primary cervical dystonia inherited as an autosomal dominant trait [181]. Those affected in the family exhibited focal cervical dystonia, occasionally with mild tremor, having its onset in early adulthood to late midlife (18 – 66 years of age). However, affectation status was not always clear cut: 5 family members were said to have 'definite' cervical dystonia, whilst 5 family members were said to have 'possible' cervical dystonia, making linkage analysis difficult [182].

CIZ1 encodes Cip1-inderacting zinc finger protein 1, a $p21^{Cip1/Waf1}$-interacting zinc finger protein expressed in the brain and involved in DNA synthesis and cell-cycle control. Functional work suggests that the p.Ser264Gly mutation may alter splicing of the gene and normal subnuclear localization of CIZ1 protein [181]. Screening in a cohort of patients with adult-onset dystonia identified 2 additional missense mutations in 3 individuals (p.Pro47Ser and p.Arg672Met), all with focal cervical dystonia developing in mid-to-late life.

Some researchers feel it may be too early to place full confidence in the association of mutations in *CIZ1* and focal cervical dystonia. Firstly, databases of normal sequence variation reveal that the *CIZ1* gene contains a large number of missense variants in supposedly healthy individuals (a third of which are predicted to be pathogenic) and, in the study by Xiao *et al.* an equal number of novel missense variants were found in controls as were found in patients. Secondly, and more importantly, segregation was not demonstrated in a second independent family for any of the purported mutations. Thirdly, there is some question regarding the quality of the exome data used in this study: coverage was only greater than twenty reads for around 60% of the exome and the number of false-positive variants appears unusually high [183]. Screening of this gene in further cohorts of cervical dystonia cases from various populations will be required to decide if these reservations are warranted or not.

3.6.5 TUBB4A Mutations (DYT 4)

DYT4 was first described in 1985 by forensic psychiatrist Neville Parker in a large family with third decade onset of autosomal dominantly inherited 'whispering dysphonia' and generalized dystonia [184]. Over twenty-five affected individuals have been reported, typically presenting with a laryngeal dysphonia progressing to a generalized dystonia and a peculiar 'hobby horse' gait [185]. Alcohol responsiveness was not uncommon, leading to severe alcohol abuse in some DYT4 patients [185]. The family is descended from an affected male who was born in 1801 in the small rural coastal town of Heacham in Norfolk and, to date, no other kindred has so far been described worldwide with a similar phenotype.

Using a combination of linkage analysis and exome sequencing, a mutation in the gene *TUBB4A* (p.Arg2Gly) has recently been identified as causal in the DYT4 kindred by two groups independently [186, 187]. *TUBB4A* encodes β-tubulin-4a, a constituent of microtubules, and the mutation results in an arginine to glycine amino-acid substitution in the key, highly conserved autoregulatory MREI (Methionine–Arginine–Glutamic acid–Isoleucine) domain of the protein. The gene is expressed throughout the brain, but at the highest levels in the cerebellum, which has been linked to the pathogenesis of dystonia [186]. The MREI tetrapeptide sequence at the start of the N-terminal domain is known to be necessary for the autoregulation of the β-tubulin mRNA transcript and separate *in vitro* studies using site directed mutagenesis have previously demonstrated that the p.Arg2Gly mutation abrogates this autoregulatory ability [188, 189]. One further possibly pathogenic variant (p.Ala271Thr) was detected during the screening of a cohort of 394 unrelated dystonia patients: the individual concerned exhibited spasmodic dysphonia with oromandibular dystonia and dyskinesia with an age at onset of 60. Her mother had been similarly affected, but was now deceased so that segregation analysis was not possible, meaning the pathogenicity of the variant is uncertain [187].

3.7 The Paroxysmal Dystonias

The paroxysmal dystonias are characterized, in theory at least, by episodes of dystonia or other dyskinesia separated periods of neurological normality. In practice, there may

be other associated neurological features besides the movement disorder and, especially in the case of *SLC2A1*-related disease, both the movement disorder and the other associated neurological symptoms may become relatively fixed with time.

3.7.1 MR1 Mutations (DYT8)

Symptoms of paroxysmal non-kinesigenic dyskinesia (PNKD) typically begin in childhood or adolescence and include dystonic and choreatic dyskinesias or ballistic movements, lasting from minutes to hours. There are often premonitory symptoms of an impending attack (paraesthesia or tension in the affected area) and sleep can prevent or abort attacks. The frequency of attacks varies between individuals, ranging from daily attacks to only a few in a lifetime. Typically, they are precipitated by alcohol, caffeine or stress, but less often by exercise, fatigue or cold [190]. Neurological examination is normal between attacks. Treatment is with carbamazepine or benezodiazepines, particularly clonazepam.

The condition results from mutations in the myofribillogenesis regulator (*MR1*) gene and is inherited as an autosomal dominant trait. Only three mutations have been described to date, all clustered in the N-terminus of the protein: p.Ala7Val, p.Ala9Val and p.Ala33Pro [191-196]. Despite this, haplotype analysis has failed to reveal a common founder, suggesting that these represent independent mutational events [194]. It is currently believed that this region of the protein may represent a mitochondrial targeting sequence [196]. Mutations may act by disrupting protein processing *in* vivo, a hypothesis that is supported by evidence from transgenic mice [197]. The function of the MR1 protein is not fully understood, but it shows close homology to glyoxalase hydroxyacylglutathione hydrolase, with is known to detoxify methylglyoxal, a compound found in coffee and alcohol and a by-product of glycolosis, thus providing a link to the standard triggers [191].

3.7.2 PRRT2 Mutations (DYT10)

Mutations in *PRRT2* cause paroxysmal kinesigenic dyskinesia (PKD), which has an estimated prevalence 1 in 150,000 individuals [198]. Affected individuals have short (seconds to minutes) and frequent (up to 100 times per day) attacks of dystonic or

choreiform movements, precipitated by sudden movements or startle. As with PKND, there is often warning of an impending attack – a so-called 'aura' – consisting of numbness or paraesthesia in the affected body part. The attacks usually begin in childhood and are highly responsive to anticonvulsant therapy, such as carbamazepine [52]. Reports of what was, in reterospect, probably this disorder have appeared in the literature from as early as 1892 and the condition was designated dystonia 10 in 1998 [199, 200]. However, the discovery of the gene underlying the condition, *PRRT2*, would wait until exome sequencing approaches could be brought to bear on the problem in 2011 [46, 47].

Most mutations are truncating and by far the most common of these is the c.649dupC mutation, but missense variants (possibly with reduced penetrance) have also been described [201-204]. The PRRT2 protein is highly expressed in the central nervous system and is probably localized to the synapse [48]. Using yeast 2-hybrid screening, it has been shown to interact with synaptosomal protein 25 (SNAP25), suggesting a role in the fusion of the synaptic vesicles to the plasma membrane [205]. Both PRRT2 and SNAP25 are highly expressed in the basal ganglia and disrupted neurotransmitter release has been suggested as a possible pathogenic mechanism in *PRRT2* mutations. Truncating mutations produce a protein lacking the transmembrane domain, and are thought to result in altered subcellular localization [47]. Missense mutations may cause loss of function or act in a dominant-negative manner [46, 206].

PRRT2-related disease is notable for its varied presentation, differing not merely between individuals carrying different mutations, but also between individuals carrying the same mutation and even between affected members within the same family [52]. As well as the classical PKD phenotype described above, infantile convulsions with choreoathetosis (ICCA) with or without PKD, benign familial infantile seizures, episodic ataxia, hemiplegic migraine and even benign paroxysmal torticollis of infancy all appear to be possible manifestations of mutations in this gene [48-51]. The common thread would appear to be the paroxysmal nature of all *PRRT2*-related disorders.

Finally, it would appear that the family used to define DYT19 also carry the common c.649dup mutation in *PRRT2*, suggesting that the initial linkage was incorrect [50].

3.7.3 SLC2A1 Mutations (DYT18)

Dominantly inherited mutations in the *SLC2A1* gene are now known to be the cause of paroxysmal exercise induced dyskinesia (PED) [207, 208]. In fact, *SLC2A1* mutations cause a wide spectrum of disease, resulting from a deficiency of the encoded glucose transporter, GLUT1. GLUT1 is the principal glucose transporter in the brain and it has generally been assumed that neuronal dysfunction arises from energy failure. Under this assumption, the prevalence of movement disorders was thought to reflect the heightened sensitivity of the basal ganglia to energy deficits [209]. However, recent studies in a mouse model of GLUT1-deficiency, which recaptures many of the features of the disease, demonstrated the preservation of aerobic metabolism, such that mechanisms other than impaired aerobic glycolysis must be responsible for the manifestations of the disorder [210]. Possible culprits include reduced transfer of lactate generated via anaerobic glycolysis from the astrocyte into the neuron and reduced anaerobic ATP formation in the astrocyte having a negative impact on glutamate-glutamine recycling, which, if proven, would place the emphasis on glial dysfunction in the pathogenesis of GLUT1-deficiency [210, 211].

At its most severe, GLUT1 deficiency can present with early onset psychomotor delay, drug resistant epilepsy, acquired microcepahly and other signs of widespread neurological dysfunction, such as spasticity, ataxia, tremor, dystonia, choreoathetosis and ballism, which may all be paroxysmal, static or static with paroxysmal worsening. From this extreme, there exists an entire spectrum of decreasing disease severity, ranging through developmental delay and movement disorders without epilepsy, right down to the mildest cases, in which only minimal phenotypic abnormalities are detectable [212, 213]. There can be considerable heterogeneity of clinical presentation even within the same family [212].

PED is, therefore, probably best regarded as one just possible manifestation of the complex and variable disorder that is GLUT1 deficiency. It is characterized by attacks

of combined chorea, athetosis, and dystonia precipitated by exercise – particularly brisk walking or running – that usually begin in childhood [207, 208]. The legs and feet are by far the most commonly affected body parts during attacks, which generally last from a few minutes to an hour [198]. Unlike PKD or PNKD, affected individuals do not report aura-like symptoms prior to an attack. The symptoms tend to improve with intravenously administered glucose and with permanent ketogenic diet, though they can become static with time [208]. Since PED is simply a manifestation of GLUT1 deficiency, it may occur alone in minimally affected patients or may be accompanied by various other manifestations of the disease, such as epilepsy, ataxia, haemolytic anaemia or spastic paraplegia. No doubt this wide variation in clinical presentation contributed to the erroneous assignation of DYT9 (paroxysmal choreoathetosis with spasticity) and DYT18 (PED with or without epilepsy) as two separate conditions; it is now known that the causative gene in both cases was, in fact, *SLC2A1* [23].

Given the variable expressivity of *SLC2A1* mutations, researchers have sought correlations between genotype and phenotype. It has been suggested that multiple exon deletion is associated with the early-onset, more severe form of the disorder, whilst missense mutations are associated with less severe cognitive deficits and a lower rate of movement disorders [212].

3.8 The Combined Dystonia Syndromes

The combined dystonia (previously 'dystonia plus') syndromes represent a heterogeneous group of diseases, where dystonia is accompanied by other neurological features but is not part of a complex neurodegenerative disorder.

3.8.1 SGCE Mutations (DYT11)

Mutations in *SGCE* cause myoclonus-dystonia, the commonest of the combined dystonia syndromes with a prevalence of about 2 per million in Europe [214]. The condition is inherited in an autosomal dominant fashion but with a significant role played by imprinting (see below). It usually begins in childhood, with a mean age of onset of ~6 years, and is characterized by brief lightning-like myoclonic jerks, most often affecting the neck, trunk and upper limbs, in combination with focal or

segmental dystonia in around two thirds of patients [215]. Myoclonus usually dominants the clinical picture and can be reliably provoked by complex motor activities, such as writing or copying a picture. There is often a dramatic improvement after alcohol, but this feature is not specific to the condition and is not present in all cases [216].

The *SGCE* gene encodes the 438-aminoacid protein ε-sarcoglycan, which contains a single transmembrane domain. Despite its ubiquitous expression pattern and close homology to α-sarcoglycan (68%), mutations in *SGCE* do not to lead to any detectable deficits in peripheral nerve or muscle function. Over 50 different mutations have been reported as causing myoclonus-dystonia, most of them located within the portion of the gene encoding the extracellular domain of the protein [217]. It is thought that mutations lead to mislocalisation of the protein from the plasma membrane to the endoplasmic reticulum and to the promotion of its degradation by the proteasome [218]. Interestingly, in humans, transcription occurs almost exclusively from the paternal allele, whilst transcription from the maternal allele is silenced by promotor methylation [219]. Because of this fact, 95% of those inheriting an *SGCE* mutation from their mother will not manifest any symptoms of the disease.

In patients exhibiting alcohol responsiveness, GABAergic drugs such as clonazepam or gamma-hydroxybutyrate often produce a pronounced but temporary improvement in the condition. However, their short duration of action can lead to overuse and addiction and this danger is compounded by apparently higher rates of psychiatric comorbidity – particularly OCD, anxiety and alcohol dependency – in *SGCE* mutation carries [220, 221]. The myoclonus seen in this condition is subcortical and does not respond to agents used to treat cortical myoclonus, such as valproate, levetiractem or piracetam [222]. Levodopa responsivemess (of both the myoclonic and dystonic elements of the condition) has been reported in two cases and some have suggested a trial of levodopa is warranted [223].

3.8.2 ATP1A3 Mutations (DYT12)

Mutations in *ATP1A3* are primarily associated with rapid onset dystonia-parkinsonism. Familial cases show autosomal dominant inheritence with reduced penetrance (~90%),

but *de novo* mutations are common so that the absence of a family history should not preclude consideration of the condition [224]. Most mutations are missense, affecting highly conserved transmembrane or N-terminus domains. The mechanism by which such mutations actually cause rapid onset dystonia parkinsonism is not fully understood. *ATP1A3* encodes the catalytic unit of a sodium pump that uses ATP hydrolysis to exchange Na+ and K+ across the cell membrane, thus maintaining the ionic gradients important for electrical excitability, neurotransmitter transport, volume regulation and other key cellular functions. Studies using cells transfected with mutant ATP1A3 have suggested mutations may reduce affinity for Na+ and that the resultant dysfunction in ion transport may impair cell viability, but the relative importance of these mechanisms to the actual pathophysiology of rapid onset dystonia-parkinsonism remains uncertain [225, 226].

Clinically, the condition is characterized by the onset, over minutes to days, of dystonia and parkinsonism, usually in the context of specific physical or psychological stressors, particularly febrile illness. After this period of rapid development of symptoms, the condition usually stabilizes and continued progression has only been reported once [227]. There is often a rostrocaudal gradient of symptoms. Despite clinical parkinsonism, dopamine transporter PET and SPECT studies reveal normal dopamine uptake [127, 228] and, unfortunately, patients with rapid-onset dystonia parkinsonism do not respond to levodopa or pallidal DBS [229, 230].

Interestingly, *de novo* heterozygous mutations in *ATP1A3* also appear to underlie a rare neuropaediatric condition termed alternating hemiplegia of childhood (AHC) [44, 231]. AHC is characterized by the onset before the age of 18 months of paroxysmal neurological events, such as hemiplegia alternating in laterality, quadriplegia, dystonic spells, oculomotor abnormalities and autonomic dysfunction, all of which abate during sleep. Non-paroxysmal manifestations develop after a few months or years of the disease, comprising developmental delay, intellectual disability of variable degree, ataxia, dysarthria, choreoathetosis, and, in some patients, pyramidal tract signs. In one study, a mutation in *ATP1A3* was demonstrated in all 24 patients with AHC [44]. A second study found *ATP1A3* mutations is just under three quarters of their cases [231].

One mutation, p.Asp801Asn, was found 30-40% of all cases, suggesting it is a mutational hotspot.

3.8.3 PRKRA Mutations (DYT16)

Mutations in the gene *PRKRA* appear to be a rare cause of autosomal recessive dystonia-parkinsonism. Camargos *et al.* identified two apparently unrelated consanguineous Brazilian families with a total of 6 individuals exhibiting young-onset, generalised dystonia or dystonia-parkinsonism [232]. Autozygosity mapping revealed a shared segment of homozygosity segregating with the disease and subsequent mutational screening of the region identified a homozygous missense mutation (p.Pro222Lys) in all six individuals [232]. *PRKRA* encodes protein kinase, interferon-inducible double-stranded RNA-dependent activator, which, in response to extracellular stress activates latent protein kinase PKR, a protein involved in signal transduction, cell differentiation, cell proliferation, antiviral response and apoptosis [233]. The Pro222Lys mutation alters a conserved amino acid between the second and third RNA-binding motifs, but the manner by which it causes disease is remains elusive. As with patients carrying mutations in *ATP1A3*, response to standard medical therapy was minimal or absent.

3.9 Dopa-Responsive Dystonia

Clinically, the dopa-responsive dystonias often present as combined or complicated dystonia syndromes, but they have been separated out in this discussion because they are bound together aetiologically by their connection to the endogenous pathway for the synthesis of dopamine, which accounts for their shared clinic characteristic of an improvement in symptoms in response to treatment with oral L-dopa. To date, three genes have been convincingly shown to cause dopa-responsive dystonia: *GCH1* (GTP Cyclohydrolase 1), *TH* (Tyrosine Hydroxylase) and *SPR* (Sepiapterin Reductase) [234-236]. All three genes encode enzymes required for the biosynthesis of dopamine (see figure 7).

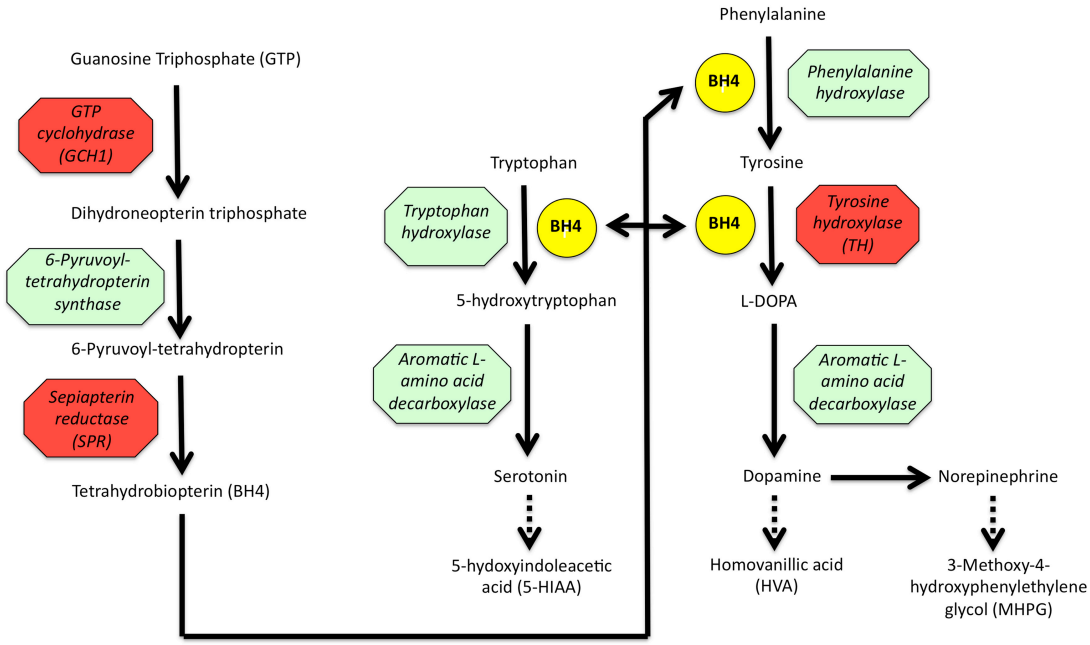

Figure 7 - Schematic illustration of the pathways for the synthesis of catecholamines and serotonin. Breakdown products are indicated by broken arrows. Enzymes in which defects cause DOPA-responsive dystonia are shown in red with the gene symbol in brackets. Tetrahydrobiopterin is a key cofactor in some enzymatic reactions.

Table 8 - CSF neurotransmitter metabolite profiles in dopa-responisve dystonia secondary to *GCH1*, *TH* and *SPR* mutations. 5-HIAA = 5-hydroxyindoleactetic acid; MHPG = 3-methoxy-4-hydroxy-phenylethylene gycol.

CSF metabolite	GCH1	TH	SPR
Biopterin	Decreased	Normal	Normal to slightly raised
Neopterin	Decreased	Normal	Increased
Homovanillic acid	Normal to mildly decreased	Decreased	Very low
5-HIAA	Normal to mildly decreased	Normal	Very low
MHPG	Normal to mildy decreased	Decreased	Decreased

3.9.1 GCH1 Mutations (DYT5a/Segawa's Disease)

Mutations in this gene account for around 60-80% of autosomal dominant dopa-responsive dystonia [237]. Penetrance is incomplete and lower for males (≈40-50%) than females (≈80%) [22]. GC*H1* encodes the enzyme GTP cyclohydrolase 1, which catalyzes the first step in the reaction of tetrahydrobiopterin neo-synthesis from GTP. Tetrahydrobiopterin is an essential cofactor for tyrosine hydroxylase, the rate-limiting enzyme for catecholamine synthesis, as well as for tryptophan hydroxylase and phenylalanine hydroxylase, which are involved in the production of serotonin. GC*H1* mutations act in a dominant negative manner, with measurable enzyme activity usually being less than 20% of normal, resulting in a relative deficiency of dopamine and serotonin [238, 239]. Diagnosis can be confirmed on the basis of reduced CSF levels of total biopterin (of which tetrahydrobiopterin is the main component) and neopterin (a product of the reaction involving GCH1) (see table 8) or an abnormal phenylalanine loading test [240, 241].

Clinical presentation is typically with lower limb dystonia and gait disturbance in the first decade of life, which occasionally leads to misdiagnosis as cerebral palsy or spastic paraparesis [242]. Diurnal fluctuation in the severity of the dystonia is common, though this may abate with age. With time there is usually gradual generalization, and parkinsonism and dystonic tremor are possible complications of the condition [243, 244]. In a subgroup of patients, the dystonia is mild, progresses only very slowly and may require no treatment [245]. Infrequently, patients may present later in life with a dystonic tremor, akinetic-rigid parkinsonism or even myoclonus dystonia [245-247]. There is some evidence that subtle neuropsychiatric features may be associated with mutations in this gene. In one recent study of 18 patients, major depressive and sleep disorders occurred in about half of those over 20 year of age, whereas obsessive–compulsive disorder was found in 25% of cases [248]. In terms of treatment, the response even to low doses of L-dopa is generally excellent (70-100% improvement in clinical symptoms), sustained and is not generally associated with the late-onset dyskinesias that often accompany prolonged use of L-dopa in other conditions, such as Parkinson's disease [249]

It should be noted that a previously-proposed novel form of dopa-responsive dystonia, DYT14, is now known to be synonymous with DYT5a. Misclassification of one patient from the 'DYT14' family had led to incorrect locus assignment on the basis of erroneous linkage. After removal of this patient, the linkage peak now included the *GCH1* gene, which was found to habour a causative deletion [22].

Finally, autosomal recessive mutations in *GCH1* have also been reported, resulting in no detectable enzyme activity in the liver. As might be expected the phenotype is generally distinct and much more severe, with complex neurological dysfunction, including developmental delay, spasticity, seizures and physiological hyperphenylalaninaemia, which is generally picked up on routine screening of newborn infants [250]. However, atypical presentations have been described with dopa-responsive dystonia [251], with a neonatal dopa-responsive extrapyramidal syndrome [252] or without hyperphenylaninaemia [253, 254].

3.9.2 TH Mutations (DYT5b)

Rarely, dopa-responsive dystonia can also be inherited as an autosomal recessive disorder associated with mutations in the gene encoding tyrosine hydroxylase itself [255]. It is sometimes referred to as DYT5b. Only about 40 cases have been reported worldwide but with over 30 different pathogenic mutations, scattered throughout the length of the gene, including its promotor regions [256].

As shown in figure 7, tyrosine hydroxylase is involved in the conversion of phenylalanine to dopamine, but it is also the the rate-limiting enzyme in the synthesis of other catecholamines, such as norepinephrine and epinephrine. The enzyme thus plays a key role in normal functioning of both dopaminergic and noradrenergic neuronal populations and, indeed, complete loss of tyrosine hydroxylase activity appears to be lethal in knock-out mice [257]. Patients with biallelic mutations in this gene tend to present with a more severe disease than is seen in those carrying heterozygous *GCH1* mutations, which reflect the more profound deficiency of dopamine and norephinephrine. Two major phenotypes can be discerned: 1) a progressive hypokinetic-rigid syndrome with dystonia beginning in the first year of life, with

significant improvement in response to L-dopa; and 2) a more complex encephalopathy, associated with variable degrees of hypokinesia, bradykinesia, hypotonia, dystonia, myoclonus, ptosis and eye movement abnormalities that has its onset in the first few months of life, is often poorly responsive to L-dopa and carries a far worse long term prognosis [249, 256]. In this latter type, autonomic functions are often disturbed leading to so-called 'lethargy-irritability crises', consisting of excessive drooling, sweating, body temperature instability and marked periods of 'pyrexia of unknown origin' [256].

In keeping with the underlying biochemical defects, CSF levels of homovanillic acid and 3-methoxy-4-hydroxyphenylethylene glycol (MHPG), the breakdown products of dopamine and norepinephrine respectively, are found to be low. CSF levels of biopterin and 5-hydroxyindoleacetic acid are normal, however, reflecting the fact that the synthesis of tetrahydrobiopterin and serotonin is unaffected by mutations in the tyrosine hydroxylase gene (see table 8 and figure 7).

3.9.3 SPR Mutations

Mutations in the gene *SPR* result in a second form of autosomal recessive dopa-responsive dystonia. Deficiency of the encoded enzyme, sepiapterin reductase, results in neurologic deterioration due to severe catecholamine and serotonin deficiencies in the central nervous system, caused by the resultant defect in tetrahydrobiopterin synthesis. As well as being a cofactor for tyrosine hydroxylase, tetrahydrobiopterin is also a required cofactor for phenylalanine hydroxylase, but patients with sepiapterin reductase deficiency do not exhibit overt hyperphenylalaninaemia as is seen in other autosomal recessive disorders of tetrahydrobiopterin synthesis as alternative enzymes are able to replace sepiaterin reductase in peripheral tissues and can thus compensate to some degree. In the brain, however, the low activity of dihydrofolate reductase means that an alternate pathway for tetrahydrobiopterin synthesis cannot complete and an intermediate, dihydrobiopterin, accumulates instead [258]. The accumulation of dihydrobiopterin is probably of pathological significance since it is a competitive inhibitor of both tyrosine and tryptophan hydroxylase, thus further compounding the deficits in catecholamine and serotonin production caused by the subnormal

tetrahydrobiopterin concentrations. Furthemore, by displacing prebound tetrahydrobiopterin, it has been suggested that dihydrobiopterin may uncouple nitric oxide synthase and cause the release of potentially apoptotic levels of superoxide and peroxynitrite, resulting in neuronal cell death [258, 259].

The complex biochemical abnormalities described above are reflected in CSF metabolite measurements which show severe decreases in homovanillic acid and 5-hydroxyindoleacetic acid in the context of a normal neopterin and a normal or slightly raised total biopterin (see table 8). The neurologic phenotype of SPR deficiency is quite variable, with milder cases being particularly susceptible to misdiagnosis as cerebral palsy [260]. Most affected individuals will display motor and speech delay, axial hypotonia, dystonia, weakness and oculogyric crises with diurnal variation and sleep benefit, whilst dysarthria, parkinsonian features, hyperreflexia, behavioural abnormalities, mental retardation, and autonomic signs are not uncommon accompaniments [260, 261]. Many patients will show considerable benefit from treatment with low-dose L-dopa, with most improvement being observed in motor and sleep symptoms. Significant cognitive problems may remain, however, and some patients have shown further benefit from the addition of 5-HTP, a serotonin precursor, leading to the suggestion that it should be trialled in all patients without complete symptom resolution on L-dopa alone [260].

3.10 Mechanistic Insights from the Monogenic Primary Dystonias

As with other disorders, it is believed an understanding of the genetic architecture of dystonia offers that best starting point for the complex task of unravelling the molecular pathophysiology of the disorder. Identification of genes involved in rarer, Mendelian forms of dystonia, along with a knowledge of their cellular function, allows us to identify key molecular pathways, which, when compromised, might contribute to the onset of dystonia. Since primary dystonia seems to result from a chronic cellular dysfunction, rather than a degenerative process, there are greater grounds for the belief that it may eventually be possible to ameliorate the disease by correcting the underlying dysfunction in these pathways. With this in mind, what do the functions of the causative genes identified so far tell us about the pathophysiology of primary dystonia?

One inescapable observation, evident from even a cursory look at the genes mentioned in the preceding sections above, is the apparent range of cellular pathways implicated in the pathogenesis of dystonia. This suggests that the phenotypic heterogeneity of the dystonias is matched by a similar heterogeneity of pathological mechanism and that dysfunction in distinct pathways or cellular components is capable of resulting in the common clinical manifestations of the disease. Nonetheless, several themes can be discerned on closer examination. Firstly, and perhaps unsurprisingly, dopamine signalling problems are clearly an important pathogenic mechanism in dystonia. All three of the genes involved in dopa-responsive dystonia (*GCH1*, *TH*, and *SPR*) are directly involved in the dopamine synthesis pathway (see figure 7); *GNAL* encodes the alpha subunit of G protein that directly couples to D1 receptors; and knock-in mouse models of the GAG deletion in *TOR1A* have been shown to exhibit neurochemical and structural changes in the dopamine pathways of the brain [262]. Regulation of gene expression and cell cycle dynamics also appear to be important: the protein products of *THAP1* and *TAF3* are both transcription factors, whilst that of *CIZ1* is a DNA replication factor that acts to transform nuclei poised to begin DNA synthesis into nuclei actively synthesising DNA [263]. With the recent identification of *TUBB4A* as the cause of whispering dysphonia, the function of several genes now point to cell structure as playing an important role in dystonia. *TUBB4A* encodes tubulin, the major component of the cytoskeleton; TorsinA appears to important for the structure of the nuclear envelope where it may form a bridge complex between it and the cytoskeleton [264] and also interacts with vimentin, a type III intermediate filament important for motility, chemotaxis, adhesion, intracellular signalling and neurite outgrowth [155]; and *SGCE*, a gene linked to myoclonus-dystonia, encodes a member of the dystrophin-glycoprotein complex that connects the cytoskeleton to the extracellular matrix in muscle and may have a role in synaptic organisation in the CNS [265]. Synaptic dysfunction is also highlighted by the recent identification of *PRRT2*, the protein product of which interacts with the SNARE protein, SNAP25 [48], in addition to evidence suggesting TorsinA may regulate vesicular traffic and, by extension, neurotransmitter turnover [266]. *ATP1A3*, the cause of rapid onset dystonia-parkinsonism and alternating hemiplegia of childhood, encodes the catalytic subunit of

a Na+/K+ pump that is expressed in neurons and cardiac cells and that is involved in maintaining ionic gradients [224].

3.11 The Heredodegenerative Dystonias

This category comprises a large number of complex neurological disorders of which dystonia can sometimes be a significant feature. However, it is important to recognise that dystonia may not be present in all cases or, more often still, may not be the dominant neurological sign within the clinical picture. Many of the conditions have a known genetic cause, but a full exposition of their clinical features and molecular basis is beyond the scope of this review. Instead, they are summarised in the table 9 by mode of inheritance.

Notably, DYT3-related dystonia belongs to this category of disease and is perhaps worth a more detailed consideration. It is primarily found in Filipino males, due to a founder mutation, and is inherited in an X-linked recessive fashion. It is due to a 2.6kb retrotransposon insertion in intron 32 of the *TAF1* (TATA box-binding protein-associated factor 1) gene [267]. The insertion appears to reduce neuron-specific expression of TAF1 and the dopamine receptor D2 gene in the caudate nucleus [267, 268]. Symptoms start as focal dystonia and progress to multifocal/generalized dystonia, sometimes with parkinsonism. Neuronal degeneration on postmortem analysis has been described in association with mutations in this gene [269, 270].

3.12 The Genetics of Sporadic Dystonia

To date, no genome-wide association study has been reported for sporadic dystonia. Previous experience in allied diseases, such as Parkinson's disease or Alzheimer's, suggests that to be successful such a study would need to be a large-scale, international effort that included thousands of case and control subjects. Such research is, of course, expensive and would require a considerable funding commitment. A second problem that a potential association study would face is defining its case cohort. Sporadic dystonia is an umbrella term, comprising multiple phenotypic subtypes that may well have distinct genetic architectures. In this respect, it is not unlike epilepsy, a condition

Table 9 - The major forms of heredodegenerative dystonia where a genetic cause has been identified, arranged by mode of inheritance. Major clinical features besides dystonia are indicated.

Disease	Gene	Major Features Besides Dystonia
Autosomal Dominant		
Dentatorubral-pallidoluysian atrophy	*ATN1*	Ataxia, chorea and cognitive decline.
Huntington's disease	*IT-15*	Chorea, depression and cognitive decline.
Huntington's disease-like 2	*JPH3*	Chorea, parkinsonism, ataxia and cognitive decline
Neuroferritinopathy	*FTL*	Chorea and parkinsonism.
SCA 2	*ATXN2*	Ataxia, ocular movement disorders, spasticity and parkinsonism
SCA 3	*ATXN3*	Ataxia, spasticity, parkinsonism and ocular movement disorders.
SCA 7	*ATXN7*	Ataxia, pigmentary macular degeneration, brisk reflexes
SCA 17	*TBP*	Chorea, parkinsonism, ataxia and cognitive decline
Autosomal Recessive		
Ataxia-telangectasia	*ATM*	Ataxia, oculomotor apraxia, telangiectasia, susceptibility to malignancy.
Choreoacanthocytosis	*VPS13A*	Chorea, orofacial dyskinesias, cognitive decline, axonal neuropathy
Ceroid-lipofuscinosis	*Multiple genes*	Visual failure, cerebral atrophy and seizures
Dystonia with brain manganese deposition	*SLC30A10*	Hepatic cirrhosis, polycythemia and hypermanganesaemia.
Fucosidosis	*FUCA1*	Mental retardation, growth retardation, dysostosis multiplex, angiokeratoma

FBOX07-associated neurodegeneration	*FBX07*	Parkinsonism
Glutaric acidaemia type 1	*GCDH*	Macrocepahaly, encephalopathic crises, axial hypotonia, parkinsonism.
Hereditary dopamine deficiency syndrome	*SLC6A3*	Pyramidal tract signs, eye movement disoreder, hyper and hypokinetic movement disorders
Kufor-Rakeb syndrome	*ATP13A2*	Parkinsonism
Infantile striatonigral degeneration	*NUP62*	Choreoathetosis, spasticity, nystagmus, developmental regression
Metachromatic leukodystrophy	*ARSA*	Mental retardation, spasticity and bulbar palsy
Mitochondrial membrane protein-associated neurodegeneration (MPAN)	*C19orf12*	Cognitive decline, prominent neuropsychiatric abnormalities, motor neuronopathy
Neimann Pick Type C and D	*NPC1/NPC2*	Ataxia, ocular motor abnormalities (vertical gaze palsy), seizures
Neurodegeneration with brain iron accumulation type 1	*PANK2*	Parkinsonism, behavioural changes, pigmentary retinopathy in about 50%
Neurodegeneration with brain iron accumulation type 2	*PLA2G6*	Parkinsonism, pyramidal signs, cognitive decline, cerebellar ataxia
Parkin-related parkinsonism	*PRKN*	Parkinsonism
Tay-Sach's disease	*HEXA*	Psychomotor regression, dementia, blindness.
Wilson's disease	*ATP7B*	Tremor, parkinsonism
X-Linked		
Dystonia-deafness syndrome	*TIMM8A*	Progressive hearing loss, spasticity, cortical blindness
Lesch-Nyhan syndrome	*HPRT*	Choreoathetosis, ballismus, cognitive/attentional deficits, self-injurious

		behaviours
Lubag disease (DYT3)	*TAF1*	Parkinsonism
Pelizaeus-Merzbacher disease	*PLP1*	Pyramidal dysfunction, cerebellar ataxia, head tremor
Retts syndrome	*MECP2*	Mental retardation, motor regression, autistic behaviours, seizures
Static encephalopathy of childhood with neurodegeneration in adulthood	*WDR45*	*De novo* inheritance pattern, global developemental delay, parkinsonism and dementia
Mitochondrial		
Leber's Optic Neuropathy	*ND1, ND4, ND6*	Bilateral or sequential visual failure
Leigh's syndrome	Multiple genes	Optic atrophy, ophthalmoplegia, ataxia, spasticity and developmental delay/regression. May also be AR inheritance.
Myoclonic Epilepsy with Ragged Red Fibres (MERRF)	Mainly *tRNA(lys)* but others reported	Epilepsy, short stature, hearing loss

for which large association studies have so far been somewhat disappointing [271]. Difficult decisions will have to be made regarding whether to lump all sporadic dystonia together and risk 'diluting' individual signals associated with particular subtypes or whether to spilt them apart and risk compromising the power of the studies to detect an association.

In the case of sporadic primary dystonia, only candidate gene association studies have yet been published. Six of these studies have examined common single nucleotide polymorphisms (SNPs) in and around the *TOR1A* gene. Two showed strong associations between individual SNPs and susceptibility for primary dystonia in populations from Iceland, southern Germany and Austria, and the United States [272-274]. However, others have shown only modest associations in Indian, Italian and northern German populations [275-277] and yet other similar studies have been able to replicate these associations [278, 279]. In contrast to *TOR1A*, no strong associations have been reported with polymorphisms in the *THAP1* gene [280, 281]. One recent study employed a haplotype tagging strategy to cover the majority of common variability in *TOR1A*, *TAF1*, *GCH1*, *THAP1*, *MR1*, *SGCE*, *ATP1A3* and *PRKRA* [282]. No association survived correction for multiple testing in this study, though 3 variants in *GCH1* did show significant association before correction and would merit follow-up in a larger case-control cohort.

3.13 Summary

In recent years, advances in sequencing technology has accelerated the pace of gene discovery in neurology and the field of dystonia has been no exception with several new genes appearing on the scene. Indeed, there is every reason to believe that the pace of discovery will continue to accelerate as sequencing costs fall and exome sequencing gives way to affordable whole-genome sequencing, allowing examination for the first time of non-coding and regulatory regions of DNA and their potential contributions to disease.

By pointing researchers in the direction of cellular pathways that are dysfunctional in familial dystonia, the discovery of new causative genes is the first step in unravelling the

complex molecular pathophysiology that underlies not just familial but also, one would hope, sporadic forms of dystonia. Current data suggest that dystonia can result from dysfunction is a wide variety of cellular pathways, which probably reflects and in part explains the wide phenotypic variety seen in the disorder. Further functional work, particularly on those new genes that have only recently been published and will require independent confirmation, may hep us identify a core set of pathogenic mechanisms in dystonia that might form the target for novel treatments. In this respect, induced pluripotent stem cell-derived neurons would offer the advantage of a more accurate model system, as well as an initial means of assessing the effect of any proposed treatments. Finally, it remains important to examine the genetics of sporadic dystonia independently by means of a large scale, genome-wide association study but, in order for this to be successful, international co-operation and a considerable funding commitment will be required.

CHAPTER 4:

Materials and Methods

4. Materials and Methods

4.1 Case Selection, Extended Phenotyping and Samples

The kindreds that I worked during my period in research were predominantly identified by Prof. Kailash Bhatia and Prof. Nicholas Wood through their extended clinical practice. In many cases, only an index case was under their direct care and the first step accomplishing the work presented in many of the following chapters was establishing contact with the rest of the family and visiting them in order to take a history, perform a neurological examination and obtain a sample for DNA extraction. Where families had become more geographically scattered and key individuals lived abroad, interviews were conducted by phone or with the aid of Skype and DNA was obtained by saliva samples traveling by post.

Samples for mutational screening of candidate variants were selected from a library of samples held at our institution with research consent. Selection criteria are given in the relevant chapters. It must be said that the selection process was somewhat of an inexact science. Unfortunately, clinical details were not recorded in a standardised fashion and, on the whole, tended to be sparse – sometimes containing no more than a single word description of the individual's clinical condition, i.e. simply "dystonia" or "PD". This meant that many samples originating from individuals that may well have exhibited a phenotype that was a good match for the index family's disease phenotype were probably overlooked as there was insufficient information recorded to determine this. Although regrettable, it is entirely understandable and perhaps inevitable given the constraints on clinicians' time.

4.2 Ethics

The ethics for all work on biological samples obtained from individuals with any form of parkinsonism were written by me and approved by the relevant ethics committees. My work, which ended up focusing predominantly on dystonia, utilised previously existing ethics that had been written by Dr Mark Edwards and approved by the relevant ethics committees at that time..

4.3 Core Genetic Methods

The following sections give details of the materials, general protocols and software used to generate the genetic data including in this work. Any significant deviations from these general protocols are indicated in the relevant chapter.

4.4 DNA Extraction

DNA was extracted preferentially from whole blood samples where possible. Where this was not possible due to difficulties in arranging to visit the individual providing the sample or because of a reluctance on the individual's part to provide a blood sample, DNA could instead be extracted from saliva. Using this later method, a DNA sample kit and consent form could be sent via the post and returned at the individual's convenience.

4.4.1 DNA Extraction from Whole Blood

Fresh blood samples were obtained by clinical researchers, such as myself, using standard venepuncture. Samples were transported in EDTA-coated bottles to the Diagnostic Genetics Laboratory at the UCL Institute of Neurology for extraction. DNA extraction was accomplished using the FlexiGene Kit (Quiagen) according to the manufacturer's instructions as briefly summarised below.

Initially, 300ul of whole blood is mixed with 750ul of FG1 lysis buffer to liberate the cell contents. The sample is then centrifuged for 20 seconds at 10,000xg and the supernatant carefully discarded to leave only the pelleted cell cellular material. Subsequently, 150ul of FG2 buffer, which contains a protease, is added to the tube containing the pellet, before vortexing the tube to homogenise its content. After a brief 5 second centrifuge step, the tube is then placed in a heating block at 65°C for 5 minutes to allow the protease to work. Once this time has elapsed, 150ul of isopropanolol is added to the tube and the tube inverted several times to cause precipitation of the DNA. The precipitated DNA is then pelleted by centrifuging for 3 minutes at 10,000xg and the resultant supernatant discarded. The pellet is then washed of impurities by the addition 150ul of 70% ethanol, vortexing briefly and then centrifugation for a further 3 minutes at 10,000xg. The resultant supernatant is once

again discarded and the pellet allowed to air dry. Finally, the extracted and purified DNA is resuspended in 200ul of FG3 buffer by vortexing for 5 seconds and heating to 65°C. The resultant suspension is assessed for quality and concentration as detailed in section 4.5 below.

4.4.2 DNA Extraction from Saliva Samples

Genomic DNA was extracted from saliva samples using the Oragene DNA kit according to the manufacturers instructions as briefly summarised below.

Firstly, the saliva sample is incubated at 50°C in an air incubator for a minimum of two hours. Subsequently, 500ul of the solution is pipetted into a clean 1.5mL microcentrifuge tube, to which 20ul of Oragene DNA purifier added. The constituents are then vortexed briefly to ensure they are fully mixed, before the tube is placed in an ice bath to incubate for a further 10 minutes. After this time had elapsed, the sample is transferred to the centrifuge and spun at 13,000 rpm for 15 minutes to pellet the cellular debris. The resultant supernatant, containing the DNA in suspension, is transferred to a clean microcentrifuge tube and the pellet discarded. In order to precipitate the DNA from suspension, 500ul of 100% ethanol is added to the supernatant, the tube inverted 10 times and then left at room temperature for 10 minutes. The precipitated DNA is then pelleted by centrifuging at 13,000 rpm for 2 minutes and the resultant supernatant removed and discarded. The pellet is then washed of impurities by adding 250ul of 70% ethanolol, allowing the sample to stand for 1 minute and then removing the ethanol, taking care not to disturb the pellet. Finally, the DNA is resuspended in 100ul of DNA buffer by vortexing and incubation at to 60°C for one hour. The resultant suspension is assessed for quality and concentration as detailed in section 4.5 below.

4.4.3 DNA Extraction from Brain Tissue

Genomic DNA was extracted from flash frozen samples of brain tissue using the Gentra Puregene Tissue DNA extraction kit (Quiagen, Hilden, Germany). Briefly, 50mg of tissue is dissected from chips of flash frozen cerebellum whilst still frozen on a sterile plate in direct contact with dry ice. The tissue is then homogenised in 3ml of cell lysis solution that had been previously dispensed into a 15ml tube. 15µl of proteinase K is added to the tube and the mixture inverted 25 times, prior to be being incubated at 55°C overnight. The following morning 15µl of RNase A solution is added to the tube before inverting the sample 25 times and incubating at 37°C for 1 hour. Subsequently, rapid cooling of the sample is achieved by placing it in ice for 3 minutes. 100µl of protein precipitation solution is then added to the mixture and the tube vortexed vigorously 20 seconds at high speed. The sample is then centrifuged for 10 minutes at 2000xg. Meanwhile, 3ml of isopropanolol is pippeted into a clean 15ml tube, to which will be added the supernatant from the previous step, taking care not to dislodge the pellet, which is subsequently discarded. The mixture of isopropanolol and supernatant mixed by inverting the tube 50 times and then centrifuged for 3 minutes at 2000xg. The resulting supernatant is carefully discarded, taking care that the pelleted DNA remains undisturbed. The DNA pellet is then washed by adding 3ml of 70% ethanol and inverting the tube several times and repelleting of the DNA was assured by a further period of centrifugation at 2000xg for 1 minute. The resultant supernatant is then carefully removed to leave only the DNA pellet in situ, which is allowed to air dry for 5 minutes. Once all of the ethanol has evaporated, 400µl of DNA hydration solution is added to the tube containing the dried pellet and the sample included at 65 °C to dissolve the DNA. The resultant suspension was then assessed for quality and concentration as detailed in section 4.5 below.

4.5 DNA Quantification

DNA quality, and for the purpose of PCR-based work, concentration was quantified by means of UV absorbance using the NanoDrop spectrophotometer. Where greater accuracy in the initial starting concentration was required, such as for next generation sequencing (NGS), DNA was quantified by means of fluorescence intensity using the Qubit 2.0 fluorometer

4.5.1 DNA Quantification by Spectrophotometry

Measurement of the concentration of DNA in a solution by spectrophotometry relies on the fact that DNA will absorb UV light at a wavelength of 260nm. By measuring the absorbance of light at this wavelength when loaded with 1ul of solution, the Nanodrop is able to calculate the concentration of DNA in the solution and the result is displayed in ng/ul. In addition, for quality control purposes, the ratios of absorbance at 260/280nm and 260/230nm are also derived. Ratios of approximately 1.8 and 2.0 – 2.2, respectively, indicate that the DNA in solution is of good quality and free of significant impurities. DNA solutions that demonstrate absorbance ratios that significantly deviate from these norms may contain impurities that interfere with downstream reactions in PCR and thus were avoided where possible.

4.5.2 DNA Quantification by Fluorescence

Measurement of the concentration of DNA in a solution by fluorescence relies on the addition of a dye, Picogreen, which binds selectively to double stranded DNA. When bound to DNA, the dye can then be excited by light at 485nm, causing it to emit fluorescence at 530nm, which is measured by the sample reader. This allows for a more accurate determination of the amount of DNA in a solution, particularly in the presence of contaminants that would otherwise influence the UV absorbance at wavelength 260, such as RNA or salt and guanidine contamination. The method is, however, more time consuming, as it requires the calibration of the fluorometer by use of standard dilutions before use, and more expensive, as it requires the use of reagents.

To prepare samples for a analysis, a mix of the reagents is made by including 199ul of Qubit buffer with 1ul of Qubit reagent (dye) for every sample that is to be analyses. 199ul of this mix is then added to a tube containing 1ul of DNA solution, vortexed, left to stand for 1 minute and then loaded on the sample reader, which completes the analysis as described above.

4.6 Primer Design and Optimisation of PCR Conditions

Primers were designed for the target DNA sequence using the free web-based software, Primer3, and subsequently checked for potential problems related to SNP-in-primer

using the latest build of the genome via Ensemble. The oligonucleotide primers, synthesised on demand by Eurofins DNA, were then hydrated with PCR-grade water to a concentration 100ng/ul for long-term storage at -80°C. Prior to use, a 500ul working stock of the primers was created at a concentration of 10ng/ul to minimise contamination and degeneration with repeated freeze-thaw cycles and use.

Primers were tested using a stock of reference genomic DNA on a variety of PCR settings to obtain optimal PCR amplification of the desired DNA sequence alone, verified by direct visualisation after gel electrophoresis (see section 4.4 below). Generally, 10ul of Roche FastStart was combined with 1ul of each primer working stock and 1ul of a stock of the desired genomic DNA at a concentration of 25ng/ul. Once the optimal conditions had been identified, they were used for amplification of all participants' DNA.

4.7 Agarose Gel Electrophoresis

Agarose gel electrophoresis was performed to verify amplification of the desired target sequence alone and to enable a qualitative visual assessment of the amount of amplified product. A stock solution of 10x tris-borate-EDTA (TBE) solution was prepared by dissolving 121.1g of Trizma base, 7.4g of ethylenediaminetetraacetic acid and 61.8g of boric acid in 1 litre of distilled water. A working solution of 1x TBE was then prepared from this stock, heated, combined with 1.5% Ultrapure Agarose (Invitrogen) and 20ul of Gel Red (Cambridge Biosciences) and allowed to set in an tray overset with combs to create wells for sample insertion. Samples were prepared by combining 3ul of PCR product with 3ul of Orange DNA loading dye (Thermo Scientific), before being pipetted into each well. The last well of each row was filled with 6ul of midrange DNA ladder (Quiagen) to permit subsequent sizing of the amplified PCR products. A current of 100mA was then applied across the gel for a period of 30 minutes. The resultant electrophoresed PCR products were visualised using a UV transluminator.

4.8 Clean-up of PCR Product for Sequencing

Prior to sequencing amplified PCR product must be purified to remove unbound primers and left-over dNTPs. Two methods were used at different stages of this work to achieve this goal: filtration and enzymatic digestion.

4.8.1 PCR Clean-up by Filtration

Using this method, 80ul of purified PCR grade water was added to the PCR product from each reaction in a 96 well plate and the entire mix transferred to an ultrafiltration plate (Millipore). The filtration plate was then placed on a vacuum under moderate negative pressure for approximately 5 minutes or until the membrane was dry. Subsequently, the PCR product trapped on the membrane was resuspended in 50ul of purified PCR-grade water by means of gentle agitation on a unheated thermo-shaker for approximately 30 minutes.

4.8.2 PCR Clean-up by Enzymatic Digestion

Using this method, 5ul of the PCR product is combined with 2ul of a prepared enzymatic clean-up solution in a fresh 96 well plate. The enzymatic clean-up solution consists of a mix of 50ul of exonuclease I (which removes single stranded DNA such as unincorporated primers) and 200ul of fast alkaline phosphatase (which eliminates unused dNTPs), diluted to 1ml by addition of 750ul of purified PCR grade water. The plate containing the mix of PCR product and enzymatic clean-up solution is then placed on a thermal cycler and run at temperature of 37°C for 30 minutes to drive enzymatic activity and then at 80°C for 15 minutes to deactivate the enzymes.

4.9 Sequencing Using BigDye Terminator v3.1

Sequencing of PCR products was performed using BigFye Terminator v3.1 chemistry (Applied Biosciences) with some adaptations made to reduce reagent cost without reducing sequencing quality. Each reaction was constituted so as to contain 0.6ul of BigDye Terminator v3.1, 2ul of 5x sequencing buffer, 0.6ul of forward or reverse primer diluted to a concentration of 10pmol/ul, purified PCR product (3.8ul if purified by filtration and 3ul if purified by enzymatic digestion) and enough PCR-grade

water to make a final reaction volume of 10ul. The reactions were placed in a 96 well plate in a thermal cycler set to run the invariable program shown in table 10.

Table 10 - Thermal cycler settings for sequencing using BigDye Terminator v3.1

Set-up of Thermal Cycler For Sequencing Using BigDye Terminator v3.1		
Denaturation	1. Rapid thermal ramp to 96°C 2. Hold at 96°C for 1 minute	
Core Sequencing Cycle	3. Rapid thermal ramp to 96°C 4. Hold at 96°C for 10 seconds 5. Rapid thermal ramp to 50°C 6. Hold at 50°C for 5 seconds 7. Rapid thermal ramp to 60°C 8. Hold at 60°C for 4 minutes	REPEAT CYCLE 25 TIMES
Reaction End	9. Rapid thermal ramp to 4°C 10. Hold at 4°C until removal for purification	

4.10 Purification of Sequencing Reactions

Unincorporated dye terminators must be completely removed prior to electrophoresis. Excess dye terminators in sequencing reactions obscure data in the early part of the sequence and can interfere with base calling. During the period of this work, two different methods were used to achieve this goal: either, passage of the reactions through an ultrafiltration plate or, alternatively, through Sephadex columns.

4.10.1 Sequencing Reaction Purification by Filtration Plate

Using this method, 80ul of purified PCR grade water was first added to the product of the sequencing reaction in each well of the plate. The content of the plate is then transferred over to a fresh, 96-well ultrafiltration plate (Millipore). The filtration plate was then placed on a vacuum under moderate negative pressure for approximately 5 minutes or until the membrane was dry. Subsequently, the PCR product trapped on the membrane was resuspended in 25ul of purified PCR-grade water by means of gentle agitation of the light-shielded ultrafiltration plate on a unheated thermo-shaker for approximately 30 minutes.

4.10.2 Sequencing Reaction Purification by Passage Through Sephadex Columns

Using this method, a hydrated solution of Sephadex was prepared by mixing 2.9g of Sephadex G-50 powder (Sigma-Aldrich) with 40ml of autoclaved water and allowing the solution to stand for 30 minutes. The solution is then briefly vortexed, before 350ul of the mixture is transferred to each well of a Corning FiltrEx 96-well plate. The plate and its contents are then centrifuged for 3 minutes at 700xg. A fresh, 96-well collection plate is then placed under the Corning FiltrEx plate and the entire volume of the sequencing reaction from each well pipetted onto the Sephadex column in the corresponding well of the Corning FiltrEx plate. The plate containing the Sephadex columns along with the underlying collection plate is then centrifuged for 5 minutes at 910xg. The wells of the collection plate will now contain the purified sequencing reactions

4.11 Electrophoretic Separation of the Sequencing Reaction Products and Sequence Analysis

Electrophoretic separation of the sequencing reaction products and reconstruction of the underlying sequence was performed using the ABI Prism 3700 DNA analyser (Applied Biosystems). The resultant electropherograms were visualised using Sequencher software (Gene Codes Corporation). This software has some ability to automatically detect variations with respect to a reference sequence, but this function is imperfect and thus all sequences were, in addition, visually inspected for additional undetected variation.

4.12 Genotyping by DNA Microarray

Data from genome-wide genotyping of SNPs by DNA microarray was used for the purpose of linkage analysis, homozygosity mapping, copy number variant (CNV) detection and association analysis. Human CytoSNP beadchips (Illumina), which genotype approximately 220,000 markers, were used to generate data for linkage analysis. For homozygosity mapping and CNV detection, where denser genotyping is desirable, the Human OmniExpress beadchip was used instead, which provides genotypes for approximately 715,000 markers spread across the genome. Finally, for

the association analysis, samples were genotyped using the custom-designed ImmunoChip. Details of the design and intended purpose of this chip can be found in the review by Cortes *et al.* (2011) [283] or the initial paper announcing its conception by Lorenz *et al.* (2003) [284].

Samples were processed, hybridised and scanned in accordance with manufacturers instructions at UCL Genomics, generating raw intensity files for further downstream use. Clustering, normalisation and calling of the resultant data was performed in-house using Genome Studio 2010 (Illumina)

4.13 Autozygosity Mapping

In order to perform autozygosity mapping, the raw intensity files for all affected and unaffected siblings available were first loaded in Genome Studio, along with the relevant manifest and cluster files for the Illumina Human OmniExpress DNA array chip. The observed genotypes for all markers in all samples were then called. The quality of the data was first verified using the appropriate metrics to ensure there had not been any chip sample failures. Provided this was not the case, the genotype calls were exported from Genome studio in the form of PED and MAP files for use with PLINK. Subsequently, PLINK was invoked and SNPs that had not genotyped in all samples were sought and removed by setting level for exclusion of a marker based on missingness to 0.1 using the *–geno* command line option. Using the newly generated clean files, PLINK was instructed to search for runs of homozygosity greater than 0.5Mb in size in all samples by using the following command.

> plink –file filename.ped filename.map –homozyg –homozyg-kb 500

All other options were left in their default settings. In order to detect runs of homozygosity, PLINK employs a windowing algorithm. That is, PLINK effectively takes a window of *x* SNPs and slides this across the genome. At each window position, the zygosity of the windowed SNPs is assessed and, after allowing for a small number of heterozygous calls, which could represent genotyping errors, a call is made as to whether the window is 'homozygous' or not. Then, for each SNP the proportion of homozygous

windows that overlap its position is calculated. This metric is used to call the segments based on a threshold for the average.

The resultant files were transferred to Excel and the results sorted first by chromosome, then by segment start position and finally by sample ID. The sorted results were then manually inspected for runs of homozygosity that were shared by all affected siblings but not present in any unaffected sibling.

4.14 Linkage Analysis

In order to perform linkage analysis, the raw intensity files for all affected and unaffected individuals available within the kindred were first loaded in Genome Studio, along with the relevant manifest and cluster files for the Illumina CytoSNP DNA array chip. The observed genotypes for all markers in all samples were then called. The quality of the data was first verified using the appropriate metrics to ensure there had not been any chip sample failures. Provided this was not the case, the genotype calls were exported from Genome studio in the form of PED and MAP files for use with PLINK. Subsequently, PLINK was invoked and SNPs that had not genotyped in most (>90%) samples or were too rare to be informative (minor allele frequency < 0.5%) were sought and removed by setting the level for exclusion of a marker based on missingness to 0.1 and that based on minor allele frequency to 0.05, using the *–geno* and *–maf* command line options. Inclusion of too many markers when performing linkage analysis can lead to inflation of LOD scores due to linkage disequilibrium. Therefore, the density of the genotyping data was reduced by invoking PLINK with the command line option *–thin*, which eliminates a proportion of SNPs in a uniform but random fashion across the genome. The proportion was set so as to leave approximately 5000 random markers spread across the genome.

The cleaned and thinned genotypying files created by PLINK were then converted to the prerequisite format required by MERLIN, the freeware program used to perform the actual linkage analysis, as per the instructions on its website. A *.model* file was created specifying the presumed disease model including mode of inheritance, estimated disease allele frequency and penetrance. For the rare Mendelian disorders

dealt with in this work, the estimated disease allele frequency was set to a low value and penetrance was set, initially at least, to 0.8. The files were first checked for formatting errors, pedigree consistency and genotyping problems by invoking PEDSTATS. Providing this did not reveal any significant problems, the files were fed to MERLIN in order to perform the linkage analysis. This linkage analysis was repeated a further two times at least with newly generated random marker sets to ensure the results were consistent.

In order to generate the genome-wide linkage plots used in this work, LOD scores were generated against chromosomal position rather than marker name and transferred directly to Excel for conversion in a line graph.

4.15 Whole Exome Sequencing

Most exome sequencing data used in this work was generated in house. Library preparation and target enrichment was performed using the Illumina TruSeq Enrichment kit, as per the manufacturer's instructions. Details of this kit's target design and its performance can be found in table 2. After library preparation, sequencing of the samples was accomplished using the Illumina HiSeq 2000.

Data for a few samples – those sequenced very early in the project prior to the acquisition of our own NGS platform – was generated offsite at either AROS Applied Biotechnology or the Beijing Genomics Institute.

4.16 In-House Bioinformatics Pipeline For Exome Sequencing Data

In order to move from raw sequencing data to variant calls, we employed a custom built in-house bioinformatics pipeline. This pipeline was largely designed and maintained by Dr Vincent Plagnol and Dr Alan Pittman. A brief description of the rational behind each step as well as some basic details of the software tools to accomplish each task is provided below.

4.16.1 Removal of Low Quality Reads

The raw sequence data generated by the Illumina HiSeq is stored in the form of a FASTQ file. This text-based file combines information on the observed sequences with the associated per base quality score, known as the Phred quality score or Q score. The Q score indicates the probability that given base is called incorrectly by the sequencer. In order to generate the Q scoring scheme, parameters relevant to the particular sequencing chemistry must be analysed in a large empirical data set of known accuracy. The resultant quality score look-up tables are used in the calculation of quality scores for *de novo* sequencing data on board the NGS platform itself.

For a mathematical point of view, Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P), by means of the equation: $Q = 10 \log_{10} P$. For example, if a Q score of 20 is assigned to a base, this is equivalent to incorrect base call probability of 1 in 100 or, stated differently, to a call accuracy for that base is 99% (see table 11). In other words, roughly 1 base in every 100bp of sequence read with a Q score of 20 will be called incorrectly. If, on the other hand, a Q score of 30 is assigned to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times or, stated differently as previously, the call accuracy for that base is 99.9%. In general, a Q score of 30 is considered a benchmark of quality in NGS. Using Illumina chemistry on the HiSeq 2000, approximately 96% of base reads are expected to have a Q score of >30, with most falling within the range of 35 to 45.

Table 11 - Phred quality score translated into probability of incorrect base call and base call accuracy.

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Initial quality control of the raw sequencing data is performed by a programme called FastQC. FastQC provides a user-friendly graphical interface that allows, among other things, to check for over-represented sequences, deviation from the expected GC content, distribution of nucleotides per read position, thus allowing a fast identification of problems that can occur during sample preparation and sequencing.

4.16.2 Alignment of Reads to the Reference Genome

One of the most important and difficult steps in the processing of raw exome sequencing data is the alignment of the reads to the reference genome. A single run on the HiSeq 2000 can produce up to 6 billion paired-end reads of approximately 100bp in length, equivalent to some 600Gb of raw data when stored in FASTQ format. Despite the rapid advancement in computer hardware, the vast size of the human genome means that mapping all of these sequences to their correct region of origin remains a computationally intensive process and a trade-off between speed and accuracy is to some extent unavoidable. Although different alignment algorithms utilise different strategies to map reads with the highest level of accuracy in the quickest time possible, the core steps in involved in alignment of any particular sequence are broadly similar. Initially, the algorithm rapidly narrows down the genomic search space by selecting out a small set of regions within the reference genome from where the observed sequence could potentially originate. This subset of 'loose matches' is then fed into more accurate, but considerably slower sub-algorithms that decide which of the loose matches is the most likely region of origin. The process is accelerated by the construction of an index of the reference genome. From a computational point of view, aligning reads to a genome can be viewed as form of approximate string matching. The goal of this process is to find a pattern (the sequence read) that matches a target in a large text or corpus (the genome), allowing for some base changes and indels. Whilst it is possible to scan the whole text for the pattern, this is inefficient; it might be considered somewhat akin to searching for a particular name in a phone book in which the names appeared in completely random order. Indexing is a technique to pre-process the search text (in our case, the reference genome) to make queries faster (in much the same way as listing the names in a phone book in alphabetical order makes

queries faster). In addition, however, indexing that can also compress the size of the text.

Within our own laboratory, Novoalign was the software tool utilised in alignment of the raw sequencing data. Novoalign first constructs an index of the reference genome to speed up the mapping process, as described above. Individual sequence reads are then mapped to the indexed reference genome by means of a full Needleman-Wunsch algorithm incorporating affine gap penalties and taking account of the Q scores for each base included in the FASTQ file. This algorithm allows for mismatches of up to 8 bases and small gaps of up to 7bp per single-ended read during mapping. With paired-end reads, such those generated by the Hiseq 2000, Novoalign first aligns the all those hits with the highest level of mappability to the reference genome (see section ... for a definition of mappability and a discussion of its relevance to analysis of NGS sequencing data). Once all the most easily mapped reads have been aligned, it scans through the remaining unmapped reads in an attempt to pair the two ends and thus assign the remaining reads, allowing for larger mismatches as it does so. Importantly, for each read that is mapped to the reference, Novoalign generates a mapping quality score. This score operates in a similar manner to the Phred quality score for single base read and reflects the probability of the correct alignment, based on the number of mismatches and gaps required to make the alignment as well as the uniqueness of the observed sequence with respect to the reference genome. This metric is used in downstream analysis to aid decisions on which reads ought to be discarded due to low mappability.

Once alignment of the raw sequencing data has been completed to the best of the algoritm's ability, the resultant data is stored in the form of a Sequence Alignment/Map (SAM) file. These text-based files stores information about each aligned read, in particular, the position on the reference contig, the orientation of the read, quality of the alignment and potential further alignment possibilities of the read. However, SAM files are large and in order to increase speed when manipulating these large files, they are often converted to BAM files by re-encoding the information in

binary format. In addition, indexing of the BAM file to produce an accompanying BAI file can further speed up operations.

4.16.3 Removal of PCR Duplicates and Non-Unique Reads

Library preparation involves a PCR step that is used for amplifying the library and adding adapters required for sequencing. This step can introduce artifacts – reads or read pairs starting at exactly the same position and having the same insert length, respectively – which might otherwise overinflate the read depth and, by extension, perceived accuracy of the underlying base calls. It is necessary therefore to identify such sequences and remove them. This task is accomplished by Picard, a set of Java-based command-line utilities that manipulate SAM files. Picard identifies read pairs with identical coordinates and removes all but the one with the highest sequence quality score. Subsequently, all sequences with a non-unique alignment – that is, reads with more than one optimal alignment to the reference genome – are discarded since, in these cases, it cannot be determined from which site the read truly originates.

4.16.4 Quality Score Recalibration

Previous studies have demonstrated that the Phred quality scores issued by the sequencing platforms can sometimes deviate from the true error rate [285]. Accurate Q scores are, however, essential for variant calling algorithms, which utilise the scores assigned to each base call over lying a potential variant to help determine the likelihood that the variant is real and thus whether it should be called or not. Q score recalibration is performed by GATK. Bases are first grouped with respect to several characteristics (e.g., raw quality, dinucleotide content). Subsequently, for each such category, the empirical mismatch rate is computed and used to correct the raw Q score.

4.16.5 Variant Calling

The procedures involved in the calling of single nucleotide variants (SNVs) and indels are different and require the use of different software tools. They are thus explained separately in the paragraphs below.

In our own pipeline, SNVs were called using the combination of SAMTools and BCFTools. Early variant or SNV calling approaches relied entirely on counting the abundance of high-quality nucleotides at a singe site. Modern algorithms, however, are capable of integrating several additional sources of information within a probabilistic framework. One important source of information such as this is the prior probability of encountering a SNV at any given position. These prior probabilities can be derived from databases of normal human sequence variation or by carrying out SNV calling in multiple individuals at the same time. Moreover, by borrowing from the techniques used for imputation in genome-wide association studies, some algorithms are even able to take into account haplotype blocks when deriving the prior probabilities for nearby SNVs. Information such as this facilitates SNV calls in medium to low coverage regions.

The BAM file is first fed to SAMTools, which collects summary metrics in the input BAM (such as the number of reads showing a variant, the base Q scores, the alignment Q scores and PCR artefact information) and computes the overall likelihood of the data being correct given each possible genotype. This information is then stored in a Variant Call Format (VCF) file. Just as in the case of SAM files, the VCF file is then usually converted to binary format (BCF) for faster downstream manipulation. The BCF file is then passed to BCFTools, which does the actual SNV calling based on these likelihoods and, once again, assigns the SNV call an overall Phred-like quality score.

In order to call small indels, a software tool called Dindel was employed. Dindel relies on the realignment of sequencing reads to candidate haplotypes in order to try and accurately detect the presence and extent of a potential indel. According to its user instructions, its workflow consists of four basic steps. Firstly, Dindel extracts all candidate indels from the read alignments stored in the BAM file and simultaneously calculates the library insert size distributions for paired-end reads. Subsequently, the candidate indels are grouped into windows of approximately 120bp. Then, for every window, Dindel generates candidate haplotypes from the candidate indels and SNPs it has detected in the BAM file and realigns the reads to these candidate haplotype. In

the final step, Dindel calls the indels based on the realigned reads and stores these calls along with an associated quality score in a modified form of the VCF file (VCF4).

It should be noted that, despite an enrichment step in library preparation designed to amplify the exome only, exome sequencing does in fact generate low level reads from non-exonic areas of the genome. SNPs and indels from these regions are processed and called as described above, though low coverage makes them significantly more error prone and they are therefore stored in a separate file. This 'genomic' file was consulted on occasion when exome sequencing failed to reveal any clear causal variant in certain kindreds. However, due to the lack of any means of predicting the functional effect of the variants contained therein and the intrinsically high error rates due to low coverage depth, it never proved helpful in elucidating the genetic cause of disease in any kindred. For the purposes of this research, therefore, it might suffice to say that, therefore, that off target reads were simply discarded.

4.16.6 Variant Annotation

Accurate annotation of SNVs and indels called from NGS data is critical as it forms the basis of the character profile for each variant against which the researcher filters the data to try and isolate the causal variant. Its central role in gene-discovery by exome sequencing is explained in section 1.5 and an overview of some of the must commonly used annotations used in the filtering process is provided in sections 1.5.1 – 1.5.4 and 2.7 – 2.7.7. The software tool used to generate automatic annotations for the variant calls in the pipeline used in our laborartory was Annovar. The level of annotation provided by Annovar can be customised by downloading the appropriate databases in a modular fashion. In our laboratory, Annovar was asked to associate, where possible, the following annotations to each variant: 1) the name of the gene in which the variant resides; 2) the ensemble transcript number and codes for the amino acid and protein substitution; 3) its effect of protein coding (synonymous, nonsynonymous, frameshift indel, nonframeshift indel, stopgain or splicing); 4) zygosity; 5) read depth; 6) variant call quality score; 7) level of segmental duplication of the gene; 8) minor allele frequency of the variant in various builds of dbSNP, 1000 Genomes, NHLBI Exome Sequencing Project and Complete Genomics 69, if previously recorded therein; 8)

conservation scores from GERP++ and PhyloP; and 9) in silico predictions of pathogenicity from SIFT, PolyPhen-2, Mutation Taster and LRT.

The resulting annotated variant list is stored in the form of a Comma-Separated Values (CSV) file with a separate column for each annotation. This file can be loaded into Excel for easy viewing and filtering by the researcher.

4.17 Targeted, High-Throughput, Next Generation Sequencing

Targeted, high-throughput, next generation sequencing was performed in house using the Illumina MiSeq (Illumina, California). Probes targeting the desired amplicons were designed using Illumina's custom software, Design Studio, and shipped around 1 month later. The process of sample preparation takes about 2 days and was completed by following the manufacturer's supplied protocol, without deviation. The protocol is lengthy and is not therefore reproduced here. A general overview of some of the key steps are provided in section 2.8 and the full protocol can be accessed via Illumina's website.

4.18 Association Analysis

Although general methodologies for the completion of association studies using fresh data would best be included in the methods section of any thesis, the association analysis presented as part of this work was somewhat unusual at it involved the merging of multiple sets of genotyping data that had been generated using different DNA arrays. This involved particular challenges that can only really be understood within the context of that particular study and so, for this reason, a description of the methods used can be found instead in the chapter concerning this work.

4.19 Generation of Organism-Wide and Brain-Region Specific Expression Data

For information on organism-wide expression of genes of interest, freely available datasets based on EST tags were obtained through UCSC Genome Browser.

For expression across various brain regions in humans, in-house datasets, originating from the work of Dr Mina Ryten and Dr Daniah Trabzuni, were used. These datasets had been generated using Affymetrix Exon 1.0 ST Arrays and brain and CNS tissue originating from 137 control individuals, collected by the Medical Research Council (MRC) Sudden Death Brain and Tissue Bank, Edinburgh, UK [286], and the Sun Health Research Institute (SHRI) an affiliate of Sun Health Corporation, USA [287]. A full description of the samples used and the methods of RNA isolation and processing can be found in Trabzuni et al., 2011 [288]. As described therein, all arrays were pre-processed using Robust Multi-array Average (RMA) quantile normalisation with GC background correction and log2 transformation in Partek's Genomics Suite v6.6 (Partek Incorporated, USA) [289, 290]. Regional differences in gene-level expression were investigated using Partek's mixed-model ANOVA with gender and batch effects (date of hybridization and brain bank) included as co-factors.

4.20 Cellular Biological Experiments

Cellular biological experiments were performed as part of the necessary supporting evidence for the publication of two new disease genes as detailed in chapters 5 and 7. Although I participated as fully as my limited skills in the techniques of cellular biology allowed, the experiments were ultimately designed and interpreted by more expert minds than my own. Moreover, the particular methods used to carry them out were peculiar to each gene. For both these reasons, I have provided the specific methods only within the chapters where they are relevant. I am indebted to colleagues in cellular biology – Prof Andrey Abramov, Dr Kira Holmström, Dr Plamena Angelova and Dr Fernando Bartolomé-Robledo – for their time, guidance and patience that was undoubtedly essential to the successful publication of these new dystonia genes.

CHAPTER 5:

Exome Sequencing in Familial Tremulous Craniocervical Dystonia

5. Exome Sequencing in Familial Tremulous Craniocervical Dystonia

5.1 Introduction

Cervical dystonia is the most common form of focal dystonia seen by neurologists [291]. Previous epidemiological studies conducted in Europe and Northern England have suggested a prevalence of 5.7 - 6.1 per 100,000 persons [292, 293]. Although lifespan is not generally reduced, individuals affected by the condition can suffer considerable physical and psychosocial distress, which has been shown to have a significant impact on their quality of life [115]. Treatment remains symptomatic, with regular injections of botulinum toxin constituting the mainstay of current medical therapy.

Genetic factors are believed to play an important role in the pathogenesis of cervical dystonia as around 10-20% of patients have one or more affected family members [294, 295]. Despite this fact, a later age at onset and characteristically reduced penetrance has made it difficult to identify kindreds of sufficient size to permit traditional linkage based approaches to gene identification. At the time this work was completed, only two genes had been conclusively shown to cause autosomal dominant primary dystonia (*TOR1A* [MIM: 605204] and *THAP1* [MIM: 609520]) [136, 157]. Even taken together, however, mutations in these genes could explain only a small fraction of familial dystonia, suggesting that there remained a number of genetic factors yet to be identified.

In the chapter, I describe my work using whole-exome sequencing in combination with linkage analysis to identify the genetic cause of dystonia in a moderately-sized kindred from the United Kingdom. This kindred appeared to exhibited autosomal dominant inheritance of primary cranio-cervical dystonia [296] and mutations in the genes *TOR1A* and *THAP1* had previously been excluded. Once, a candidate causal variant had been identified, sanger sequencing of the exon containing the mutation was undertaken in a large number of dystonia samples followed by next-generation targeted sequencing of the whole gene to provide a comprehensive genetic screening of phenotypically similar cases.

5.2 Subjects and Methods

5.2.1 Clinical Details of the Index Family

All samples were collected with the written consent of participants and formal ethical approval by the relevant research ethics committee. All living individuals from the index family were re-examined and videoed as part of this study and the two now deceased individuals had been examined and videoed as part of a previous study [296]. The family pedigree is shown in figure 9.

All family members shown are over 25 years of age. All definitely affected family members exhibited tremulous cervical dystonia with a variable degree of associated upper limb dystonic tremor on examination. In addition, family member II-7 and III-7 had laryngeal involvement and family member II-4 had both laryngeal involvement and blephrospasm. Age at onset ranged from 19 - 39, with most having onset in the last few years of their 4th decade. One family member, II-1, had additional neurological signs on examination: he exhibited mild truncal ataxia, dysarthria and mild cognitive impairment, all dating from an episode of Wernicke-Korsakoff's encephalopathy 6 years previously. His family confirmed that he had consumed alcohol excessively for much of his life prior to that episode. There was no evidence of dystonia or any other neurological signs in any definitely unaffected individual. The affectation status of one individual, represented by a circle with a question mark in the centre (III-2 on the family pedigree), was uncertain. She described neck pain with a tight, pulling sensation on the left-hand side. Examination revealed a subtle left sided torticollis, but no tremor. As onset for all but one member of the family affected by the disease had been in the late 30s and she was currently 46, it was felt possible that these symptoms and examination findings may represent an early stage or a *forme fruste* of the condition and so for the purpose of linkage analysis her affectation status was set to unknown.

Genomic DNA extracted from whole blood was available for 15 individuals for analysis. DNA from individual III-8 and III-10 was only obtained at a late stage and was not available for linkage, but was used for segregation analysis.

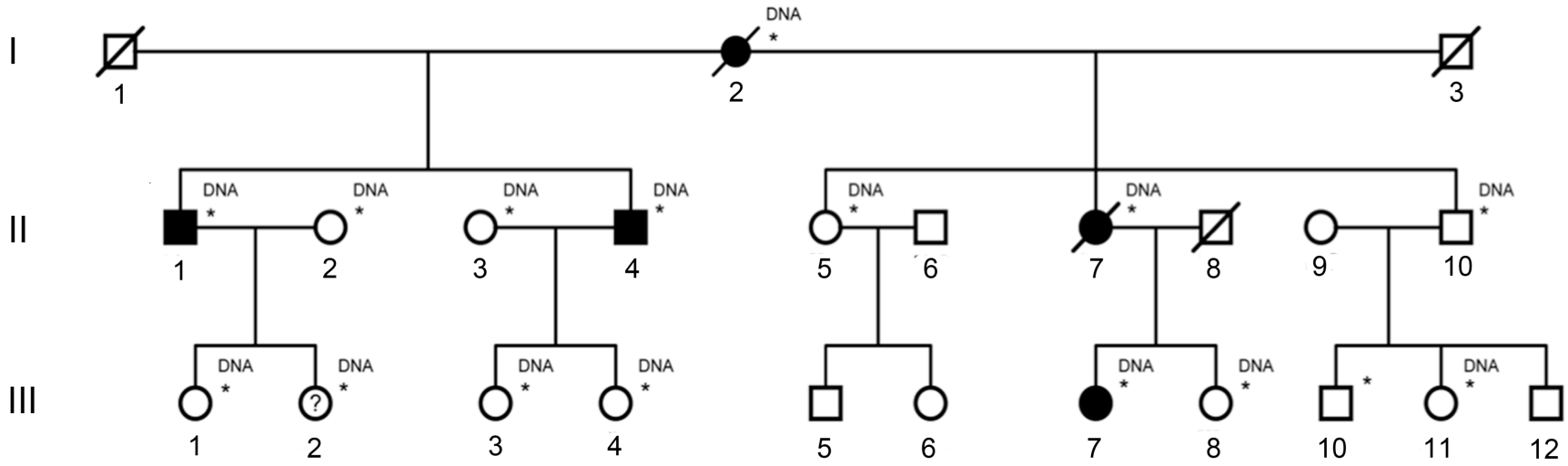

Figure 9 - Pedigree showing structure of the index family. Each generation in indicated on the left by means of a Roman numeral and each individual in that generation in indicated behind their genetic symbol in Arabic numerals. Individuals marked with an asterisk were clinically examined, videoed and reviewed by a movement disorders specialist. Family members that were felt to be definitely affected are represented by filled symbols, whilst family members that were felt to be definitely unaffected are represented by empty symbols. The affectation status of family member III-2, represented by a circle with a question mark in the centre, was uncertain. DNA was available as marked.

5.2.2 Linkage Analysis

Linkage anlaysis was performed in the standard manner as described in the section 4.14. For the parametric and non-parametric analysis, logarithm of odds (LOD) scores were calculated under the model of an autosomal dominant disease with a penetrance of 70, 80, 90 and 100%. Maximum score were produced with a penetrance of 80%

5.2.3 Exome Sequencing

3ug of genomic DNA from the two most distantly-related, definitely affected family members (individuals II-1 and III-7) was used to perform exome sequencing as per section 4.15

5.2.4 Targeted NGS Sequencing of the ANO3 Gene

Targeted high-throughput sequencing of the *ANO3* gene in 188 dystonia proband samples was performed as described in section 4.17 and 2.8. Custom oligonucleotides were designed to target all 27 exons of *ANO3* (including both untranslated regions). At least 25 intronic bases were included either side of each exon.

5.2.6 Expression Profiling of the ANO3 Gene

Expression profiling data was collated for *ANO3* as described in section 4.19

5.2.7 Ca^{2+} Imaging in Patient Fibroblasts

Fibroblasts were obtained from skin biopsy, after signed consent, from a patient carrying the (c.1470G>C; p.Trp490Cys) mutation in *ANO3*. Age and passage-matched controls were selected from in-house cell lines. The fibroblasts were cultured in Dulbecco's Modified Eagle Medium Glutamax supplemented with 10% (v/v) heat–inactivated fetal bovine serum and 1% Penicillin Strepomycin. They were maintained at 37°C in a humidified atmosphere of 5% CO_2 and 95% air.

Cytoplasmic Ca^{2+} concentration ($[Ca^{2+}]_c$) was measured using fura-2 [297], following stimulation of cells with a variety of agonists to raise $[Ca^{2+}]_c$. Cells were loaded for 30 min at room temperature with 5μM fura-2 AM and 0.005% Pluronic in a HEPES-buffered salt solution (HBSS) composed (mM): 156 NaCl, 3 KCl, 2$MgSO_4$, 1.25

KH_2PO_4, 2 $CaCl_2$, 10 glucose and 10 HEPES, pH adjusted to 7.35 with NaOH. Ca^{2+}-free medium contained 0.5mM EGTA. All analysed areas were chosen at random and three independent experiments were performed for each condition. The number of cells analysed for each set of experiments is indicated in the results section below.

Fluorescence measurements were obtained on an epifluorescence inverted microscope equipped with a 20x fluorite objective. $[Ca^{2+}]_c$ was monitored in single cells using excitation light provided by a Xenon arc lamp, the beam passing monochromator at 340 and 380 nm (Cairn Research, Kent, UK). Emitted fluorescence light was reflected through a 515 nm long-pass filter to a cooled CCD camera (Retiga, QImaging, Canada) and digitised to 12 bit resolution. All imaging data were collected and analysed using software from Andor (Belfast, UK). The fura-2 data have not been calibrated in terms of $[Ca^{2+}]_c$ because of the uncertainty arising from the use of different calibration techniques.

ATP (100μM) was used to stimulate $[Ca^{2+}]_c$ signals in fibroblasts via purinoceptors and release calcium from the endoplasmic reticulum via IP3 receptors. 50mM KCl was used induce depolarisation of the plasma membrane and open voltage gated calcium channels. Thapsigargin (1μM) in Ca^{2+}-free medium (plus 0.5 mM EGTA) was used to induce release of calcium from reticulum to cytosol and thus to estimate the size of the reticular Ca^{2+}-pool.

5.3 Results

5.3.1 Linkage Analysis

Setting the penetrance in the parametric model to 80% produced the highest linkage score. With the penetrance set at this value, 5 linkage peaks were observed with an identical maximum LOD score of 2.01 on chromosome 4, 5, 6, 7 and 11, as shown in the plot overleaf (figure 10).

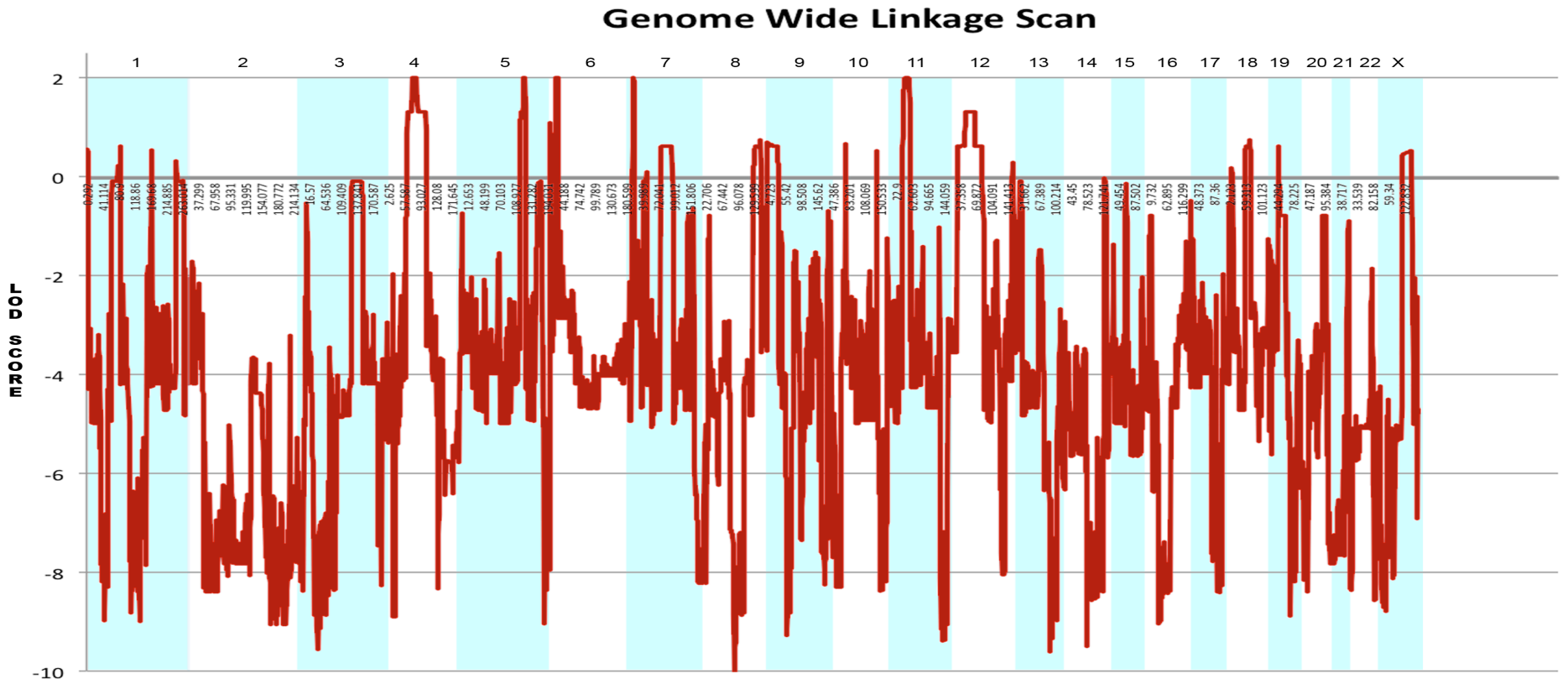

Figure 10 – Genome-wide linkage scan showing five peaks with an identical LOD score of 2.01 on chromosomes 4, 5, 6, 7 and 11. The chromosome is indicated across the across the top of the plot; chromosomal distance in centiMorgans is indicated on the x-axis; and the LOD score is indicated on the y-axis.

Details of the size and gene content was for each of these five peaks with the highest LOD score obtained from Ensemble BioMart using the hg19 build of the human genome and are summarised in table 12 below.

Table 12 - Summary of position, size and genetic content of 5 regions with highest LOD score of 2.01. The region size was calculated from the position of the first marker with a LOD score of greater than -2 to the position of the last marker to have a LOD score of greater than -2 for each peak in turn. Chr. = chromosome; Mb = megabases

Chr.	Start Position (hg19)	Stop Position (hg19)	Region Size (Mb)	Known Genes	Novel Genes	Putative Genes
4	41,749,265	101,131,209	59.4	514	238	30
5	110,825,128	118,963,199	8.1	66	45	3
6	9,275,355	14,322,665	5.0	42	28	6
7	4,883,121	10,991,660	6.1	69	37	9
11	24,192,528	42,730,635	18.5	110	76	14

5.3.2 Exome Sequencing

For the two exomes sequenced, 57,506,202 and 30,644,686 unique reads per exome were generated, translating to a total variant count of 20,935 and 17,024, respectively. Using the CCDS hg19 definition of the exome, coverage was 90% and 83% at least 2 reads and the mean read depth across the exome was 46 and 45 reads, respectively.

5.3.3 Initial Variant Filtration and Analysis

Variant analysis for exome sequencing data was based on the assumption that the mutation causing this uncommon, heritable form of the disease in this family would not be present in the general population at an appreciable frequency. In order to maximize the chances of isolating the causal variant and minimize the chances of error in assignment, two different strategies were employed to select candidate causal variants. The first strategy involved selecting only those variants that were present in the exome data of both affected family members for analysis. Homozygous variants, synonymous variants and variants recorded in dbSNP135 were initially removed. We

then filtered out any variant present at a global minor allele frequency of ≥1% in a range of publically available databases of sequence variation (1000 Genomes, Complete Genomic 69 Database and NHLBI Exome Sequencing Project database) as well as those found in 2 or more our own in-house exomes from individuals with unrelated diseases (*n*=200). Finally, variants within the regions under the linkage peaks with the highest LOD scores (on chromosomes 4, 5, 6, 7 and 11; see table 12 for definition of regions) were validated by Sanger sequencing in the forward and reverse directions using BigDye Terminator v3.1 chemistry and the Applied Biosystems 3130XL Genetic Analyser (Life Technologies, Carlsbad, CA) and checked for segregation. This strategy revealed three potentially pathogenic variants in the genes *TBC1D7*, *PPM1K*, and *ANO3*.

In order to compensate for any unequal coverage between the two exomes, a second strategy was employed in which all variants from the exome with the best coverage were first filtered as above to produce a list of every non-synonymous SNV or frameshift and non-frameshift indel that is not recorded in dbSNP135 or not recorded in public databases of sequence variation at a global minor allele frequency ≥ 1%. We then discarded any variants that were not in areas covered by the linkage peaks with the highest LOD score (see table 12). This strategy revealed variants in three further genes *FAM13A*, *TMEM232*, *LMPK1*. For each of these variants in turn, we then visually inspected the data for the other exome to ensure that it had been covered. The region containing the variant in *FAM13A* had been covered at >50x read depth and was not present in the other exome and so was discarded. However, the regions containing the other two variants had not been adequately covered in the other exome. The exon of the gene containing the variant was there Sanger sequenced and, if the variant was found to present, it was checked for segregation in the rest of the family.

The details of all the candidate variants revealed by both strategies, as well as the the results of the segregation analyses, are summarised in table 13 and discussed in the following sections in order of chromosomal location.

Table 13 - Summary of all previously unreported, potentially pathogenic variants within the regions of linkage detected within the exome with the best coverage. Only one variant in the gene *ANO3* is seen to segregate with disease in the index family. Chr = chromosome.

Chr	Gene Symbol (Transcript)	Variant	Present in other exome	Validated	Segregation	Comment
4	*PPM1K* (NM_152542.3)	c.77G>A; p.Arg26His	Yes	Yes	No	Definitely unaffected individuals III-1 (age 42) and III-8 (age 32) carry the variant
4	*FAM13A* (NM_014883.3)	c.1146_1149del; p.Asn383fs	No (50x coverage)	Yes	No	Definitely affected individual II-1 does not carry the variant
5	*TMEM232* (NM_001039763.3)	c.1111-1112del; p.Tyr371fs	Not covered	Yes	No	Definitely unaffected individual II-10 (age 54) carries the variant
6	*TBC1D7* (NM_016495.4)	c.166-167del; p.Arg56fs	Yes	Yes	No	Definitely unaffected individual II-5 (age 61) and III-8 (age 32) carry the variant
7	*LIMK1* (NM_002314.3)	c.1827G>C; p.Met609Ile	Not Covered	Yes	No	Private change to individual III-7
11	*ANO3* (NM_031418.2)	c.1480A>T; p.Arg494Trp	Yes	Yes	Yes	Present in all definitely affected; not present in any definitely unaffected

.

5.3.4 Segregation Analysis of Candidate Variants in Index Family

The first candidate variant, identified by the first strategy, affected the gene *PPMK1* (Protein Phosphatase, Mg^{2+}/Mn^{2+} Dependent, 1K; MIM: 611065) on chromosome 4. The encoded protein, essential for cell survival and development, is targeted to the mitochondria where it plays a key role in regulation of the mitochondrial permeability transition pore. Biallelic mutations in this gene are a cause of the childhood-onset metabolic disorder, maple syrup urine disease [298]. The heterozygous variant observed (c.77G>A) affected exon 2 of the gene, resulting in an arginine to histidine substitution at amino acid position 26. It is not predicted to be damaging by MutationTaster, SIFT or PolyPhen2 and is found at a low frequency African American samples (2/4404) in the NHLBI exome data. Most importantly, it did not segregate with disease in the index family, in that definitely unaffected individuals III-1, who is 42 years of age, and III-8, who is 32 years of age, both carry the variant.

The second candidate variant affected the gene *FAM13A* (Family With Sequence Similarity 13, Member A; MIM: 613299), also on chromosome 4. Little is known about the function of this gene, though it is believed to have GTPase activator activity [299]. Polymorphisms in this gene have been associated with chronic obstructive airways disease and lung cancer [299, 300]. The heterozygous variant observed (c.1146_1149del) affected exon 2 of the gene, resulting in a frameshift at 383. It was identified by the second strategy in the best covered exome, but was not present despite high read depth in the second exome (>50x; confirmed by Sanger Sequencing) and so was not considered further.

The third variant, affecting the gene *TMEM232* (Transmembrane Protein 232; OMIM: no entry) on chromosome 5, was also identified by the second strategy using the best covered exome only. Inspection of the exome sequencing data for the second exome revealed that he region of the gene harbouring the variant had not been covered. Subsequent Sanger sequencing demonstrated the variant was indeed present in this exome too. Nothing is known about the function of the protein. The variant detected (c.1111-1112del) causes a frameshift in exon 10 of the protein at amino acid 371, predicted to lead to nonsense-mediated decay. The affected amino acid is lies towards

the end of the second of two predicted transmembrane helices of 21 amino acids in length, according to Uniprot. The variant did not segregate with disease in the index family in that definitely unaffected individual II-10 (age 54) carries the variant.

The fourth candidate variant was a heterozygous frame-shift deletion in exon 2 (c.166-167del; p.Trp56*fs*) of the gene *TBC1D7* (TBC1 Domain Family, Member 7; MIM: 612655) on chromosome 6. The protein product of the gene appears to interact with the tuberous sclerosis TSC1-TSC2 complex to regulate of cellular growth and differentiation, probably through positive regulation of the mTOR-signaling pathway [301]. Homozygous truncating mutations in this gene have been shown to cause macrocephaly and intellectual disability [302, 303]. No individual affected by this disorder or, more importantly, any of their heterozygous parents have been reported to demonstrate dystonia, making the heterozygous frameshift mutation detected here an unlikely cause of cervical dystonia in our family. Indeed, it failed to fully segregate in that individual II-5, who is unaffected at age 61, and individual III-8, who is unaffected at age 32, exhibit the deletion.

The fifth candidate variant, a missense mutation (c.1827G>C; p.Met609Ile) in the gene *LIMK1* (LIM domain kinase 1; MIM: 601329), located on chromosome 7, was also identified by the second strategy using the best covered exome only. However, subsequent Sanger sequencing revealed that this change was a private mutation, which was not found in any other member of the index family, and so the variant was excluded from further consideration.

The final candidate variant, a missense mutation in exon 15 (c.1480A>T; p.Arg494Trp) of the gene *ANO3* (Anoctamin 3; MIM: 610110) on chromosome 11, segregated perfectly with the disease status in definitely affected and unaffected individuals (see figure 11). In addition, the individual of uncertain affectation status was also seen to carry the variant. The mutation occurred at a base that was highly conserved between species (see figure 12), resulting in a change from arginine to tryptophan at position 494 of the protein, and was predicted to be damaging by MutationTaster, SIFT, and PolyPhen2.

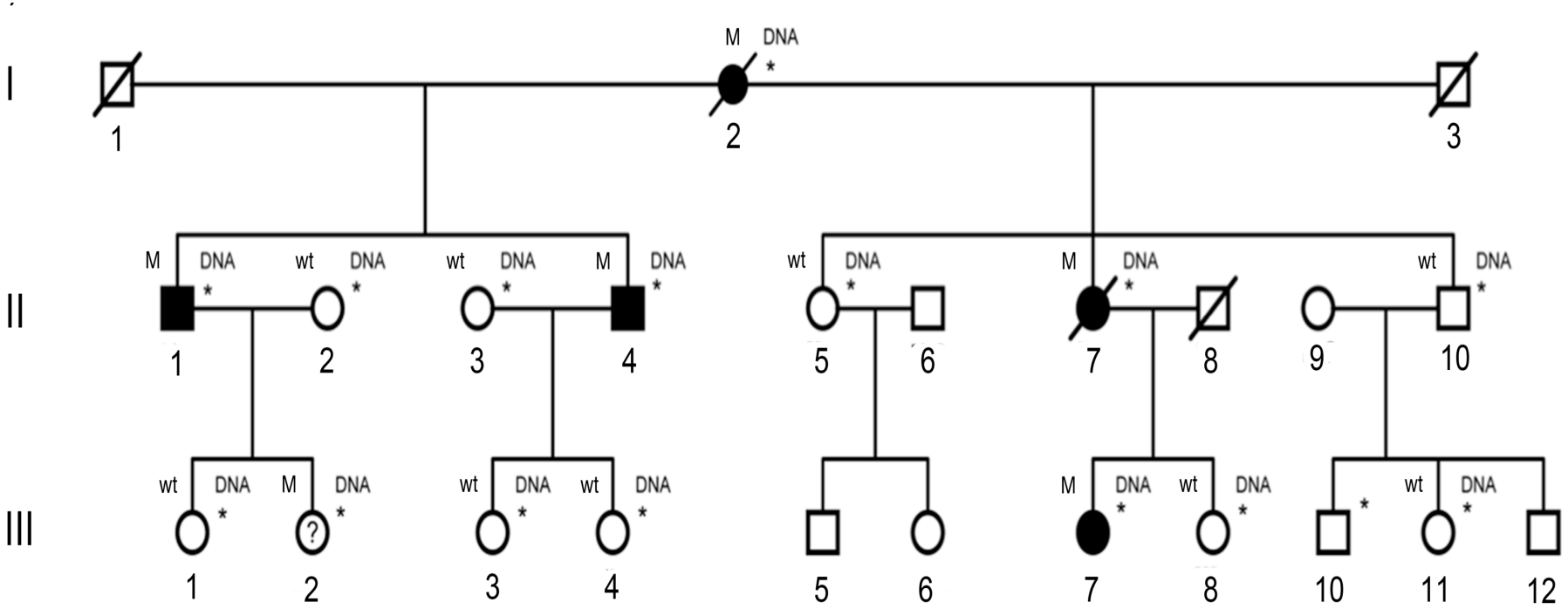

Figure 11 – Pedigree showing structure of the index family with sequencing findings for the *ANO3* c.1480A>T (p.Arg494Trp) mutation indicated above and to the left of each symbol: wt = homozygous wildtype alleles; M = heterozygous mutation carrier. Definitely affected family members are represented by filled symbols and definitely unaffected family members by empty symbols. The affectation status of family member III-2, represented by a circle with a question mark in the centre, was uncertain. DNA was available as marked. Individuals marked with an asterisk were clinically examined

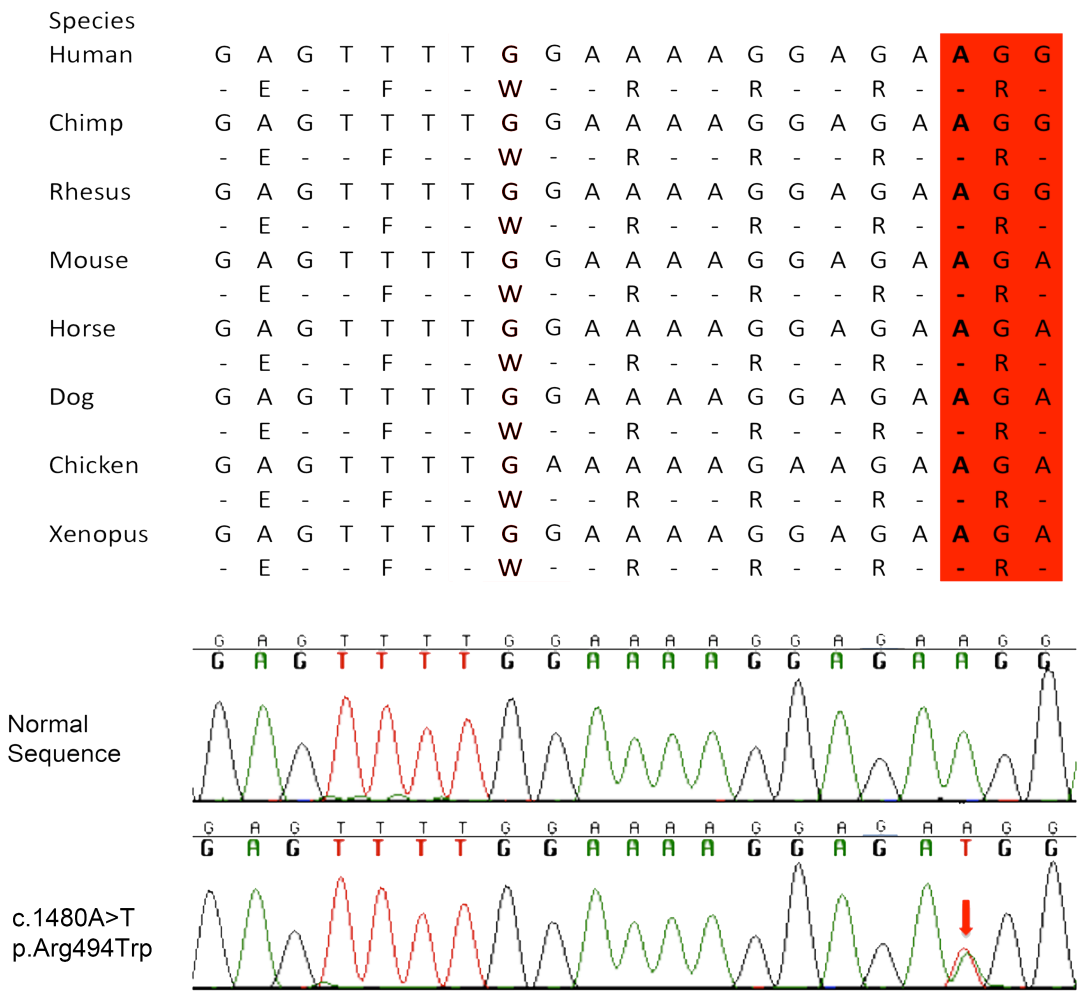

Figure 12 - Diagram showing complete conservation of protein sequence and almost complete conservation of amino acid sequence across species in the region of exon 15 of *ANO3* where the disease-segregating mutation (c.1480A>T; p.Arg494Trp) was found in the index family (affected base shown in bold and codon in red). Below, aligned electropherograms showing normal and mutated sequences, with the mutation indicated by means of a red arrow.

5.3.5 Sanger Sequencing of Exon 15 of ANO3 in a Cohort of Phenotypically Similar Dystonia Cases

Based on the data so far, I took our best candidate variant in exon 15 of *ANO3* and Sanger sequenced the exon in a selection of phenotypically similar cases. As an additional check, I also sequenced the exon containing the variant of the gene *TMEM232*, which had failed to segregate solely because of the presence of the variant in a single unaffected individual in the same selection of cases. This was done to account for the possibility of a reduced penetrance, which appears to be common in

dystonia. DNA samples for sequencing were obtained from an in-house library of previously donated samples from individuals who had given research consent. Samples were selected on the basis that the accompanying clinical description suggested cervical dystonia and/or dystonic upper limb tremor. Both familial (*n*=137) and sporadic (*n*=247) cases were selected for inclusion. Samples that were known to have previously tested positive for *TOR1A* or *THAP1* were excluded. We also included a small number of samples from individuals where the primary clinical impression had been of familial essential tremor or myoclonus dystonia (provided they tested negative for mutations in the only known connected to this disorder, *SGCE* [MIM: 604149]) as it was felt these clinical phenotypes might easily be confused with upper limb dystonic tremor and jerky cervical dystonia, respectively [304]. A total of 384 samples were screened.

Analysis of the sequence traces for these 384 individuals revealed no potentially pathogenic variants in exon 10 of *TMEM232*. In exon 15 of *ANO3*, however, we found a second heterozygous missense mutation (c.1470G>C; p.Trp490Cys) in the same highly conserved DNA sequence in which the mutation in the index family was located (see Figure 13). This second mutation was also predicted to be universally damaging by SIFT, PolyPhen2 and MutationTaster. It was not present in the data from NHLBI Sequencing Project (4090 ± 627 samples of European ancestry at a read depth of 80 ± 40), the 1000 Genomes Project or our own in-house exome data (*n*=200). The phenotype of the individual (IV-2 in figure 14) in which the mutation was found was almost identical to that for the initial index family; she had tremulous, somewhat jerky craniocervical dystonia with laryngeal involvement and a dystonic tremor of the upper limbs. Onset was in the early teens and her brother and father were also affected suggesting autosomal dominant inheritance. DNA samples were obtained from all available members of the family, which confirmed that both the father (III-1) and brother (IV-1) carried the variant, whilst the father's unaffected brother (III-3) did not. Interestingly, her paternal great-grandmother (I-1), but not her paternal grandmother (II-1), was also reported to be have been affected by head tremor. Although both individuals were by that time deceased and could not be examined, this suggests the possibility that penetrance of this mutation may not be complete.

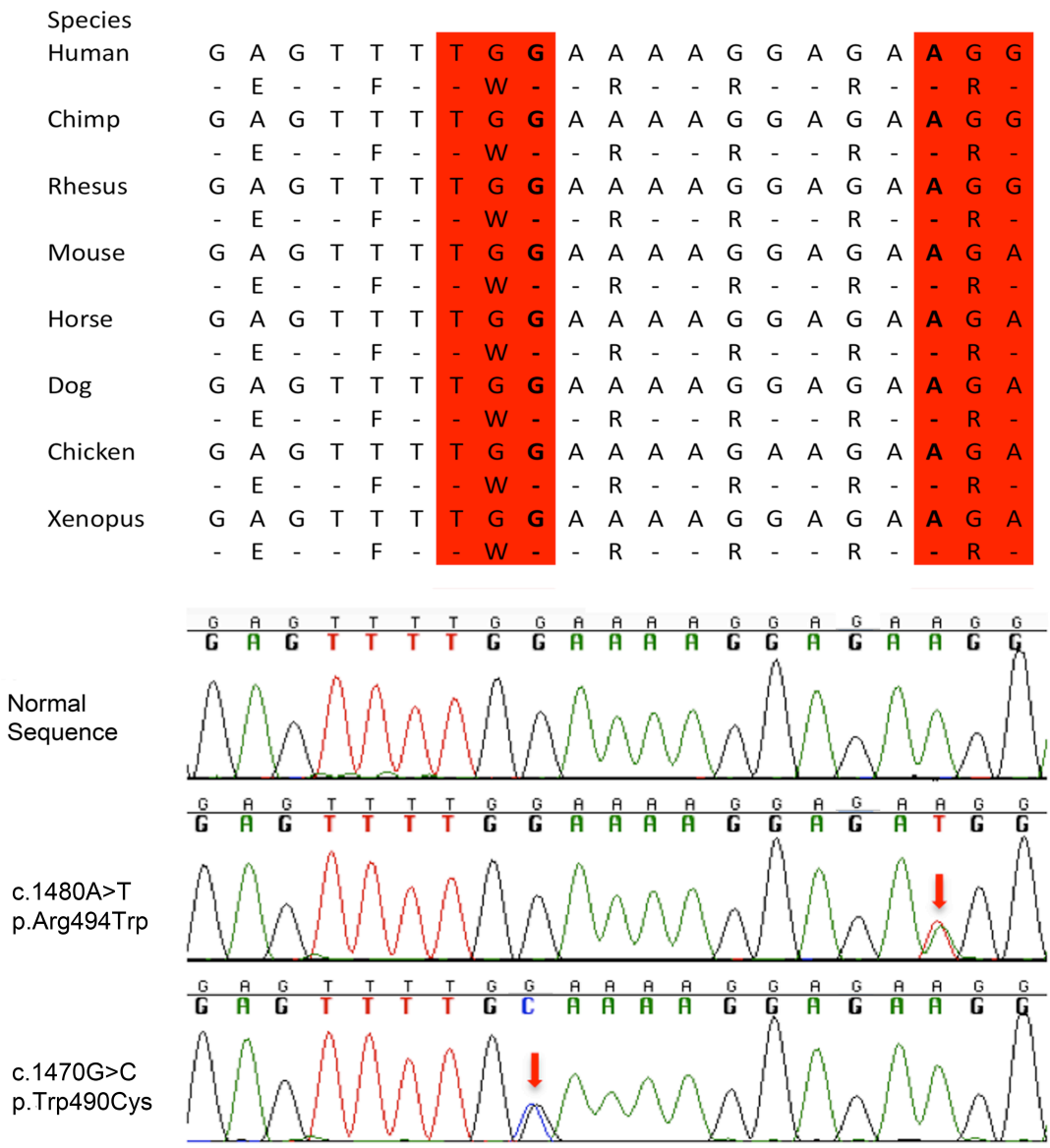

Figure 13 – Updated figure showing complete conservation of protein sequence and almost complete conservation of amino acid sequence across species in the region of exon 15 of *ANO3* in which disease-segregating mutations where found in both the index family (c.1480A>T; p.Arg494Trp) and in a second family with autosomal dominant tremulous cervical dystonia (c.1470G>C; p.Trp490Cys). Affected bases are shown in bold and affected codons in red. Below, aligned electropherograms showing normal and mutated sequences, with mutations indicated by red arrows

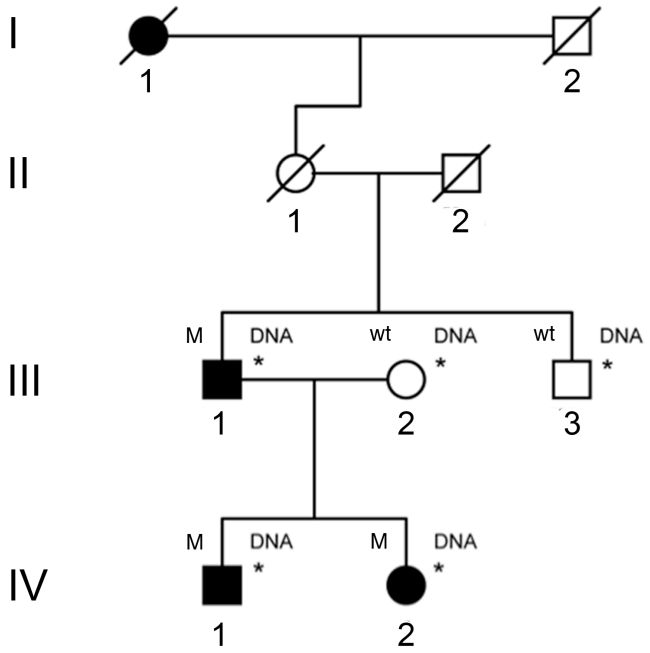

Figure 14 - Structure of the second phenotypically similar family, carrying a second, different mutation in the same exon of *ANO3* (c.1470G>C; p.Trp490Cys). DNA was available as marked. Sequencing findings for the mutation are indicated above and to the left of each symbol: wt = homozygous wildtype alleles; M = heterozygous mutation carrier. Individuals marked with an asterisk were clinically examined

5.3.6 Targeted NGS Sequencing of the Whole ANO3 Gene

Based on the finding of a second mutation, we performed targeted high-throughput sequencing of the *ANO3* gene in 188 samples as per section 4.17 A total of 110 familial and 78 sporadic samples, selected as described in section 5.3.4, were screened.

Targeted NGS sequencing identified 4 putative pathogenic variants: 3 missense variants in exons 2, 21 and 25 and a variant in the 5' UTR (summarised in table 14 and shown on figure 18). None of these variants were seen in the publically available datasets of the NHLBI Exome Sequencing Project, 1000 Genomes or within our own in-house exomes. In-silico predictions of their pathogenicity using MutationTaster, SIFT and PolyPhen2 were, however, contradictory (see table 14 for individual results). Clinically, 2 of the four individuals had cervical dystonia and 3 of the 4 individuals had upper limb tremor. The individual carrying a mutation in exon 21 of *ANO3*

(c.2053A>G; p.Ser685Gly) had a clear autosomal dominant history of cervical, laryngeal and upper limb tremulous dystonia (see figure 15). She (II-1), her mother (I-2) and her son (III-1) had all developed symptoms in their first decade of life. I was able to obtain DNA samples from her mother and son: both individuals were heterozygous for the c.2053A>G mutation. Her father was homozygous for the normal allele. She also reported one unaffected sister and one sister who developed laryngeal dystonia in her late twenties. These individuals currently reside outside of the UK and I was only able to obtain DNA from the affected sister, who was, however, heterozygous for the c.2053A>G mutation.

Unfortunately, I was unable to obtain DNA for segregation for the variants in the UTR, exon 2 or exon 25, due either to social circumstances or because the affected relatives were deceased or did not wish to take part in the study.

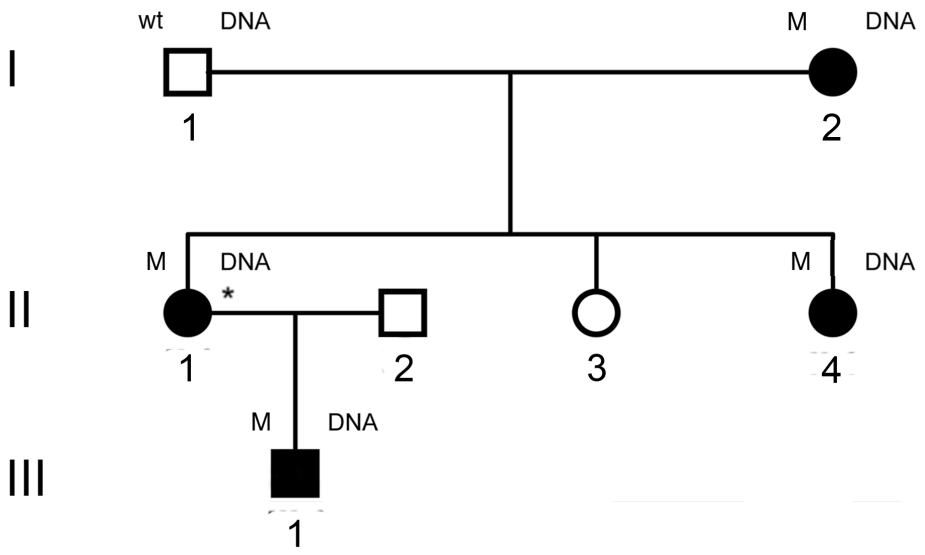

Figure 15 – Structure of the third family with phenotypically similar disease, carrying a third disease-segregating mutation in exon 21 of *ANO3* (c.2052A>G; p.Ser685Gly). DNA was available as marked. Sequencing findings for the mutation are indicated above and to the left of each symbol: wt = homozygous wildtype alleles; M = heterozygous mutation carrier. Individuals marked with asterisk were clinically examined

Table 14 - Additional variants in *ANO3* identified by high-throughput targeted sequencing of all 27 exons of the *ANO3* gene with *in silico* predictions of pathogenicity and brief clinical description of the cases. None of these variants were in the publically available datasets of the NHLBI Exome Sequencing Project, 1000 Genomes or within our own in-house exomes.

Location within transcript (NM_031418.2)	cDNA change	Protein change	MutationTaster Prediction	SIFT Prediction	Polyphen2 Prediction	Clinical phenotype of case
Exon 2	c.161C>T	p.Thr54Ile	Polymorphism	Tolerated	Benign	Diagnosed as 'familial essential tremor'; patient not contactable.
Exon 21	c.2053A>G	p.Ser685Gly	Disease causing	Tolerated	Benign	Early onset (first decade of life) autosomal dominant cervical dystonia, dystonic tremor of the upper limbs and laryngeal dystonia. Mother, one sister and son also affected. Sister affected later and by laryngeal dystonia only.
Exon 25	c.2586G>T	p.Lys862Asn	Disease causing	Tolerated	Benign	Cervical dystonia and oromandibular dystonia; deceased father was also affected
5' UTR	c.-190C>T	-	Disease causing	No prediction	No prediction	Cervical dystonia and upper limb tremor since late teens; diagnosed as myoclonus dystonia but *SGCE* gene testing negative. No family history on maternal side but father not seen since birth.

5.3.7 Regional Expression Profiling of ANO3 in the Human Brain

We then investigated the expression of the *ANO3* gene in brain and CNS tissues using publically available and in-house datasets, as described in section 4.19. Figure 16A shows the distribution of *ANO3* mRNA expression at the gene level for the following CNS regions: putamen (PUTM, n=121), frontal cortex (FCTX, n=122), temporal cortex (TCTX, n=114), hippocampus (HIPP, n=114), cervical spinal cord (SPCO, 13), substantia nigra (SNIG, n=96), hypothalamus (HYPO, n=13), medulla (specifically inferior olivary nucleus , MEDU, n=109), intralobular white matter (WHMT, n=120), thalamus (THAL, n=107) and cerebellar cortex (CRBL, n=129). This demonstrated significant regional differences in *ANO3* mRNA expression with an 5.3-fold difference (p value $<1.0 \times 10^{-45}$) between the putamen, the highest *ANO3* expressing region, and frontal cortex, the region with the second highest expression, whilst there is a 70 fold difference in expression between the putamen and the cerebellum, the region with the lowest expression (p value $< 1.0 \times 10^{-45}$). These findings are consistent with those of Kang et al., 2011 and Johnson et al., 2009 [305, 306], which demonstrate increasing expression of *ANO3* mRNA during the course of human brain development from early mid-fetal development (13< Age <16 post-conception weeks) to adolescence (12< Age <20 years) particularly within the striatum, but also the neocortex, hippocampus and amygdala (figure 16B).

5.3.8 Examination of the Effects of ANO3 Mutations on Cell Signalling by Use of Patient Derived Fibroblasts

ANO3 belongs to a family of genes that are thought to encode ion channels, more specifically Ca^{2+}-activated Cl^- channels [307]. Therefore, in order to investigate the influence of the *ANO3* mutations on calcium homeostasis, cytoplasmic Ca^{2+} concentration ($[Ca^{2+}]_c$) was measured using fura-2, following treatment of cells with ATP, potassium chloride (KCl) and thapsigargin. In the text below, *n* indicates the number of cells analysed for each set of experiments, which were universally carried out in triplicate.

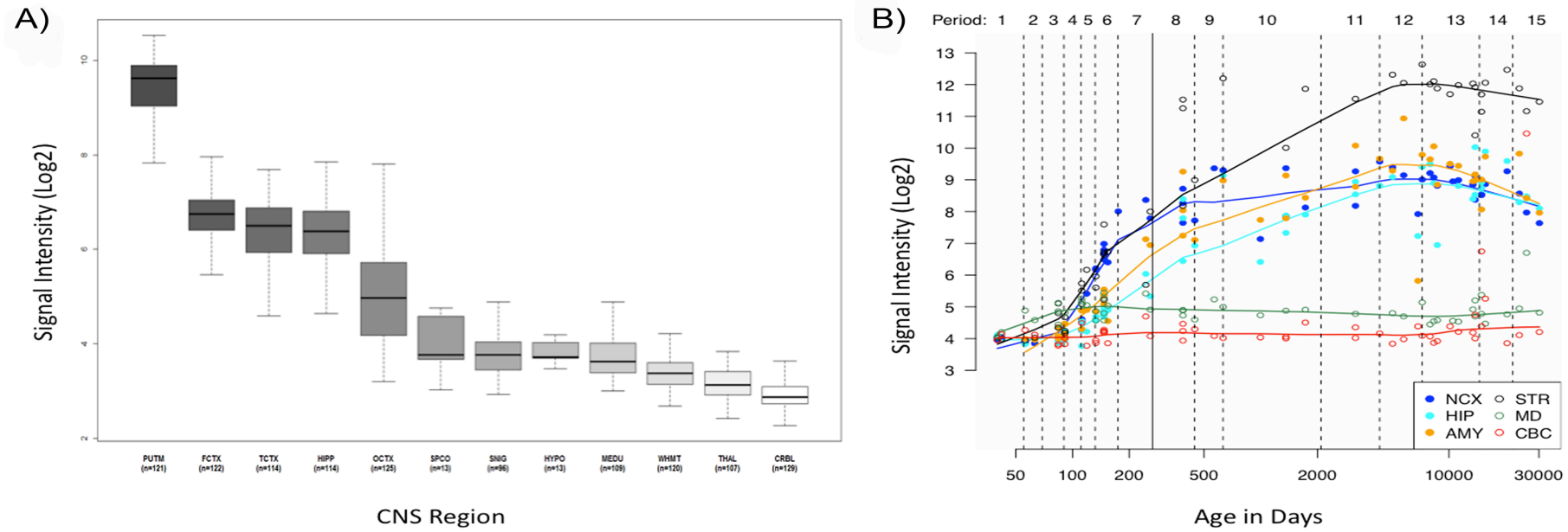

Figure 16 - A) Box plot of mRNA expression levels for *ANO3* in 12 CNS regions, based on exon array experiments and plotted on a log2 scale (y axis). This plot shows significant variation in *ANO3* transcript expression across the 12 CNS regions analysed (see main text for list of regions and abbreviations used). *ANO3* mRNA expression is significantly higher in PUTM as compared to all other brain regions. Whiskers extend from the box to 1.5 times the inter-quartile range. B) Graph to show mRNA expression levels for ANO3 in 6 brain regions during the course of human brain development, based on exon array experiments and plotted on a log2 scale [305, 306]. The brain regions analysed are the striatum (STR), amygdala (AMY), neocortex (NCX), hippocampus (HIP), mediodorsal nucleus of the thalamus (MD) and cerebellar cortex (CBC). This plot shows increasing expression of *ANO3* mRNA during human brain development from the early mid-fetal period to adolescence, particularly in the striatum.

Initially, ATP (100μM) was used to stimulate $[Ca^{2+}]_c$ signals in fibroblasts via purinoceptors and release calcium from endoplasmic reticulum via IP3 receptors. 50mM KCl was then applied to induce depolarisation of the plasma membrane and open voltage gated calcium channels. Fibroblasts from the patient carrying the c.1470G>C; p.Trp490Cys mutation showed significantly reduced ATP-induced calcium signal (figure 17A-B; n=41; $p<0.05$) when compared to control cells. Importantly in this respect, application of 50mM KCl induced a similar rise in $[Ca^{2+}]_c$ in both control fibroblasts (Ctrl-1 and Ctrl-2; n=56 and 33, respectively) and in fibroblasts with the c.1470G>C; p.Trp490Cys mutation in *ANO3* (figure 17B). Thus, the mutation in *ANO3* appeared to result in a reduced calcium signal in response to ATP that would most eloquently be explained by a smaller calcium pool within the endoplasmic reticulum. The lack of significant difference in response of control and mutation-carrying cells to depolarisation of the plasma membrane by KCl suggests that the mutation does not modulate voltage-gated calcium channels or impair active Ca^{2+} transport (mechanism of calcium removal from cytosol).

To confirm these findings, we performed further experiments using thapisgargin (1μM) in Ca^{2+}-free medium (plus 0.5 mM EGTA). Addition of thapsigargin, which is an inhibitor of the sarcoplasmic endoplasmic reticulum calcium ATPase (SERCA), induces a release of calcium from the reticulum to the cytosol and can be used to estimate the reticular Ca^{2+}-pool. Adding Ca^{2+} at the end of the experiment stimulates elevation of $[Ca^{2+}]_c$ in fibroblasts (figure 17C) due to the opening of store operated calcium channels. We found that the calcium signal in response to thapsigargin in fibroblasts carrying the c.1470G>C; p.Trp490Cys mutation was significantly smaller (n=26; $p<0.001$; Figure 17C and 17D) when compared to control fibroblasts Ctrl-1 (n=19) and Ctrl-2 (n=22). This strongly suggests that the thapsigargin-sensitive Ca^{2+} pool in the endoplasmic reticulum of mutation-carrying fibroblasts cell is significantly smaller compared to control cells. Stimulation of the store-operated Ca^{2+} channels induced similar elevation of $[Ca^{2+}]_c$ in control and mutated fibroblasts (figure 17C and 17D).

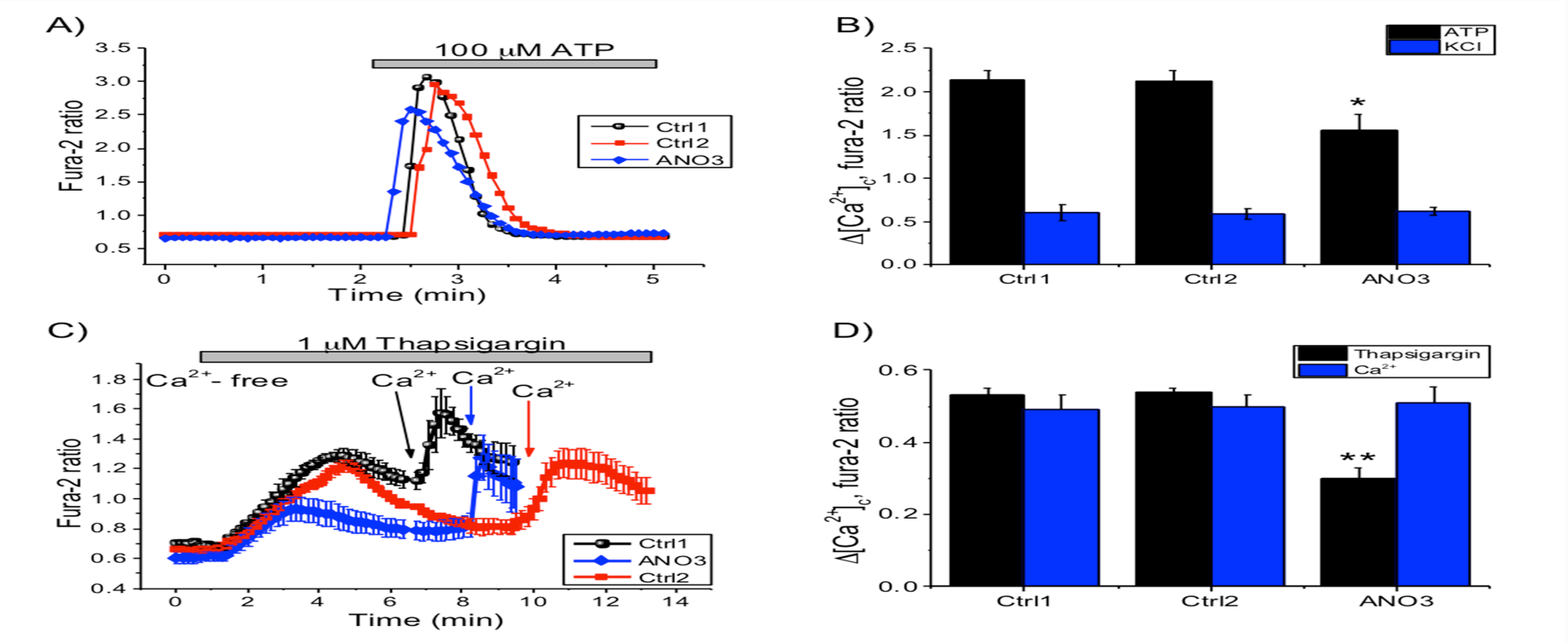

Figure 17 - Graphical Summary of Fibroblast Functional Studies. A) Typical trace of cytoplasmic Ca^{2+} concentrations as measured by fura-2 ($[Ca^{2+}]_c$) in control (Ctrl 1 and Ctrl 2) and mutation-bearing (*ANO3*, c.1470G>C; p.Trp490Cys) fibroblasts in response to application of 100 µM ATP. B) Histogram showing a significantly decreased change in the cytoplasmic Ca^{2+} concentration in response to 100 µM ATP (black bars) in mutation-bearing fibroblasts, whilst the change in the cytoplasmic Ca^{2+} concentration in response to 50mM KCl was unchanged (blue bars). Error bars represent standard error of the mean and the asterisk indicates $p<0.05$. C) Mean trace of $[Ca^{2+}]_c$ in response to thapsigargin (1 µM), and subsequent Ca^{2+} challenge (arrows; 2 mM). Error bars represent standard error of the mean. D) Histograms demonstrating a significant difference in ER calcium pool in *ANO3* mutant cells in response to thapsigargin (black bars) but no changes in the activation of store operated calcium channels as a response to the subsequent calcium challenge (blue bars) compared to controls. Error bars represent standard error of the mean and the double asterisks indicates $p<0.001$.

5.4 Discussion

In a moderately-sized kindred from the UK with apparently autosomal dominant inheritance of cranio-cervical dystonia and dystonic tremor, I performed linkage analysis and exome sequencing to identify candidate causal variants. Based on this approach, the top candidate was a missense variant in exon 15 of *ANO3* (c.1480A>T; p.Arg494Trp). Further sequencing of this exon of *ANO3* revealed a second small family with an almost identical phenotype who haboured a second missense variant 10 bases upstream (c.1470G>C; p.Trp490Cys) that segregated with the disease in the family members available for testing. Neither variant in exon 15 was seen in the publically available datasets of the NHLBI Exome Sequencing Project, 1000 Genomes or within our own in-house exomes and both were predicted to be deleterious by SIFT, Polyphen2 and MutationTaster. Subsequent targeted high-throughput sequencing of the entire gene in 188 individuals revealed 3 further coding variants and 1 variant in the 5' UTR in individuals with tremulous cervical dystonia and/or upper limb tremor. Although these variants were also absent in the above databases, predictions of their pathogenicity were contradictory and further functional work or mutational screening will be required to firmly establish their link to dystonia. However, at least one of these additional variants (c.2053A>G; p.Ser685Gly) appeared to segregate with disease in the family members available for testing, suggesting it may well be pathogenic.

ANO3 encodes a protein called anoctamin 3, about which little is yet known. It belongs to a family of genes (*ANO1-10*) that appear to be closely related in sequence and topology, but with distinct expression patterns [308, 309]. Members of the family are found throughout the eukaryotes, including mammals, flies, worms, plants, protozoa and yeast, but are best represented in higher vertebrates who possess the most members [310]. Moreover, many of these genes have been linked to disease, suggesting that they play an important role within their specific tissue types. For instance, *ANO1* (MIM: 610108) has been linked to cancer [311], mutations in *ANO5* (MIM: 608662) are linked to several forms of muscular dystrophy [312, 313], mutations in *ANO10* (MIM 613726) are linked to autosomal recessive spinocerebellar ataxia [314], and mutations in *ANO6* (MIM: 608663) are linked to Scott syndrome, a rare bleeding disorder [315].

ANO1 and *ANO2* (MIM:610109), the best studied members of the family, encode proteins that function as Ca^{2+} activated chloride channels (CaCCs) [316]. It remains an open question as to whether anoctamin 3 functions in the same manner. Hydropathy analysis suggests a similar topology, with eight hydrophobic helices that are likely to be transmembrane domains and cytosolic N- and C-termini (see figure 18), but more recent work has suggested that anoctamin 3 may in fact be targeted to the endoplasmic reticulum, rather than the cell surface, like anoctamins 1 and 2 [307]. CaCCs are, nonetheless, known to have a role in the modulation of neuronal excitability [317, 318] and, in view of our data showing very high expression of *ANO3* in the striatum, it is plausible to hypothesize that mutations in this gene might lead to abnormal excitability of striatal neurons, which then manifests itself clinically in unwanted dystonic movements. In this regard, it is interesting to note that the two mutations that we found in exon 15 lead to amino acid changes within a predicted cytosolic loop of the protein that some have suggested may function as the Ca^{2+} sensor [308]. Indeed, mutations in the homologous loop of *ANO3*'s sister gene, *ANO2*, have recently been shown to alter the voltage dependence of channel activation [319]. Our own data from patient fibroblasts carrying a mutation in exon 15 of *ANO3* confirm abnormalities in Ca^{2+} signaling. The lack of a difference in Ca^{2+} response to KCl, which is expected to open ion channels in the plasmalemma, in the context of a significantly reduced response to two agents that are known to cause calcium release from the endoplasmic reticulum (ATP and, more specifically, thapsigargin), suggests a potential defect in endoplasmic reticulum related Ca^{2+} handling in these mutation-bearing fibroblasts.

A third mutation (c.2053A>G, p.Ser685Gly) was found in the loop between the 5th and 6th transmembrane domains of the protein (see figure 18). In anoctamin 1, it is thought that this loop may form a critical component of the channel pore [320], though there remains debate about whether this loop is extracellular, re-entrant, or cytosolic [321]. If anoctamin 3 were also proven to function as an ion channel, this work might suggest a possible mechanism by which an amino acid substitution in this loop could confer pathogenicity.

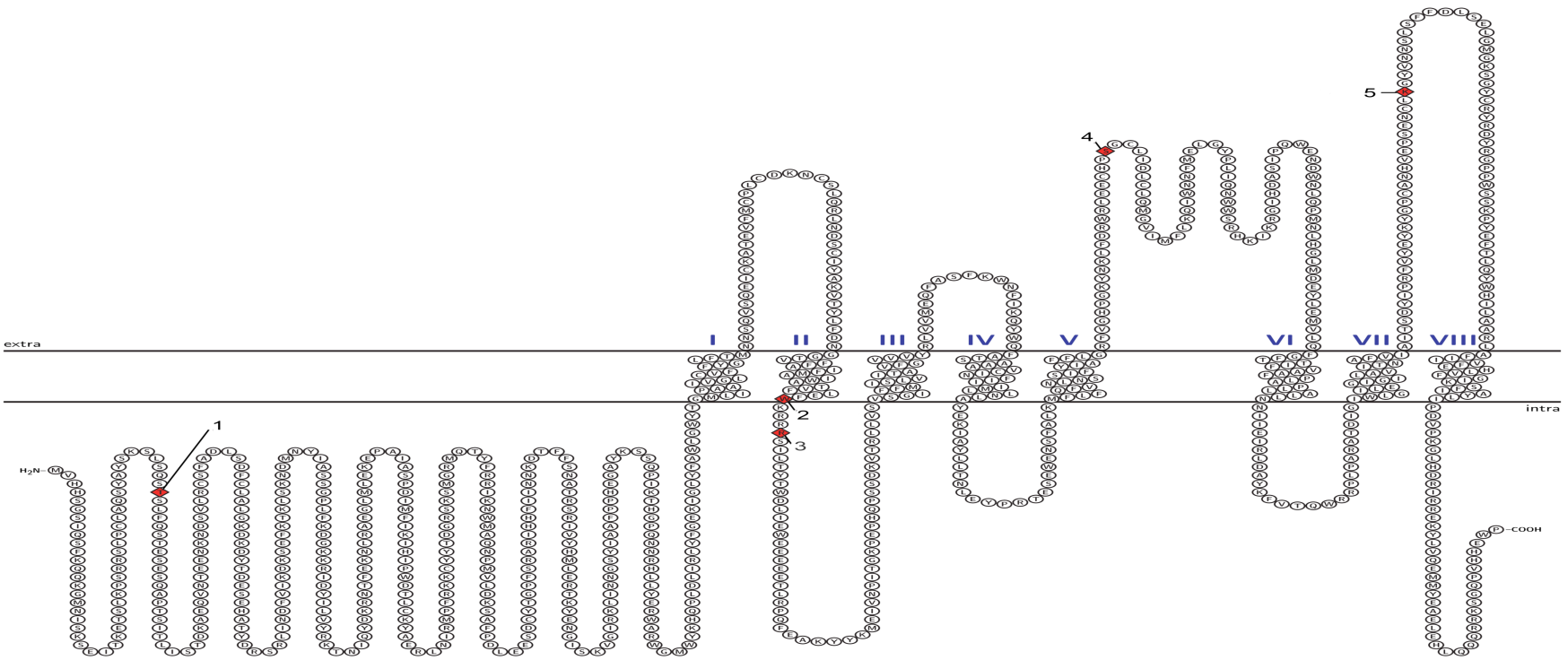

Figure 18 - Predicted topology of anoctamin 3, showing 8 hydrophobic transmembrame domains and cytosolic n- and c-termini. The loop between domains II and III may function as a Ca^{2+} sensor. The loop between domains V and VI may form a channel pore and there is debate over whether this loop may be re-entrant or even internal [321]. Amino acids highlighted in red show the position and nature of mutations found in this study. Key: 1 = c.161C>T, p.Thr54Ile; 2 = c.1470G>C, p.Trp490Cys; 3= c.1480A>T, p.Arg494Trp; 4 = c.2053A>G, p.Ser685Gly; and 5 = c.2586G>T, p.Lys862Asn.

Although further functional work will be required to establish the mechanism by which mutations in *ANO3* might lead to dystonia, the implication of a transmembrane ion channel in the pathogenesis of this condition represents a completely fresh avenue of inquiry for future research in this field and, importantly, raises the possibility that pharmaceutical agents targeted at compensating for aberrant channel function could potentially be beneficial in the treatment of a subset of dystonia patients. For instance, CaCCs can be blocked in vitro by nifulmic acid, by tamoxifen and, to a lesser extent and in a less specific manner, by fluoxetine [318]. Finally, it will be important to carry out further genetic screening of phenotypically similar cases in this and other populations in order to establish the prevalence of mutations in this gene as a cause of autosomal dominant cervical dystonia, dystonic head tremor and/or upper limb dystonic tremor.

CHAPTER 6:

Exome Sequencing in Late-Onset Tremulous Cervical Dystonia

6. Exome Sequencing in a Late-Onset Tremulous Cervical Dystonia

6.1 Introduction

As part of the work leading to the publication of *ANO3*, we sought to identify as many families as possible from our clinics and database records with tremulous cervical dystonia in the hope that they might also harbour a mutation in this gene. In the process, we identified a family with a similar phenotype of tremulous cervical dystonia that appeared to be inherited with a high degree of penetrance. Unlike individuals with *ANO3* mutations, however, who mostly developed dystonia in their teens and always before the age of 40, individuals in this family developed the first signs of dystonia in late adulthood, typically in their 7th decade. Targeted high-throughput sequencing failed to identify a mutation in *ANO3* in the family, suggesting that it had a different genetic basis.

The structure of the family (see figure 19) makes it at once attractive, whilst simultaneously presenting significant problems for the genetic researcher. The family was particularly large by Western standards, with the oldest surviving generation consisting of a sibship of 15 individuals. Of these, only 1 had passed away (reported to be unaffected), 1 lived elsewhere in the world (reported to be affected) and 1 could not be contacted (reported to be unaffected). Of the remaining 12, 7 appeared to be affected by dystonia, suggesting an autosomal dominant inheritance pattern with almost complete age-dependent penetrance. Most of these individuals had married and had children, but only 1 individual in the subsequent generation had so far developed very mild cervical dystonia. However, the typical onset age of the condition in the members of oldest surviving generation meant that their children had not yet reached the age where they might be expected to manifest the condition if they were actually carrying the causative mutation. As a result, an individual in the this generation with no clinical signs of dystonia when examined for this study could just as equally: 1) not have inherited the causative gene; or 2) have inherited the causative gene but not yet manifested the disease as they had yet to attain the typical age of onset. By extension, for all but the single individual with signs of dystonia in this generation, it was not

possible to assign affectation status with any hope of accuracy, rendering informative linkage analysis an impossibility. Thus, it was recognised from the outset that any attempt to 'solve' the family would be mired in diffiuculty because no variants could be excluded on the basis of genomic location in an area of negative linkage.

Nonetheless, despite these challenges, the large size of the first generation and high number of affected individuals proved sufficiently tempting and the family was selected for further clinical review and exome sequencing.

6.2 Subjects, Materials and Methods

6.2.1 Clinical Details of the Index Family

Participants were drawn a family of English ancestry. The pedigree is shown in figure 19. The inheritance pattern was felt to be suggestive of a late-onset, autosomal dominant condition with high age related penetrancy. Both males and females were affected, making sex linked inheritance unlikely. The ratio of reportedly affected to unaffected individuals in the oldest surviving generation was 3:2, much more suggestive of a dominant rather than recessive trait, where the ratio would be expected to be 1:3. At least one example of vertical transmission was evident (between individuals II-4 and III-11) and the family also reported that individual I-2 may have been affected (which, if this was the case, would also rule out mitochondrial inheritance).

The clinical characteristics of affected individuals were almost unvariable. In summary, all individuals exhibited a tremulous cervical dystonia, often with associated hand and laryngeal tremor, leading to voice tremor. Onset was generally in the 60s and was gradual. Though the severity of the dystonia had worsened with time, generalisation was not seen and the dystonia remained confined to the craniocervical and brachial regions.

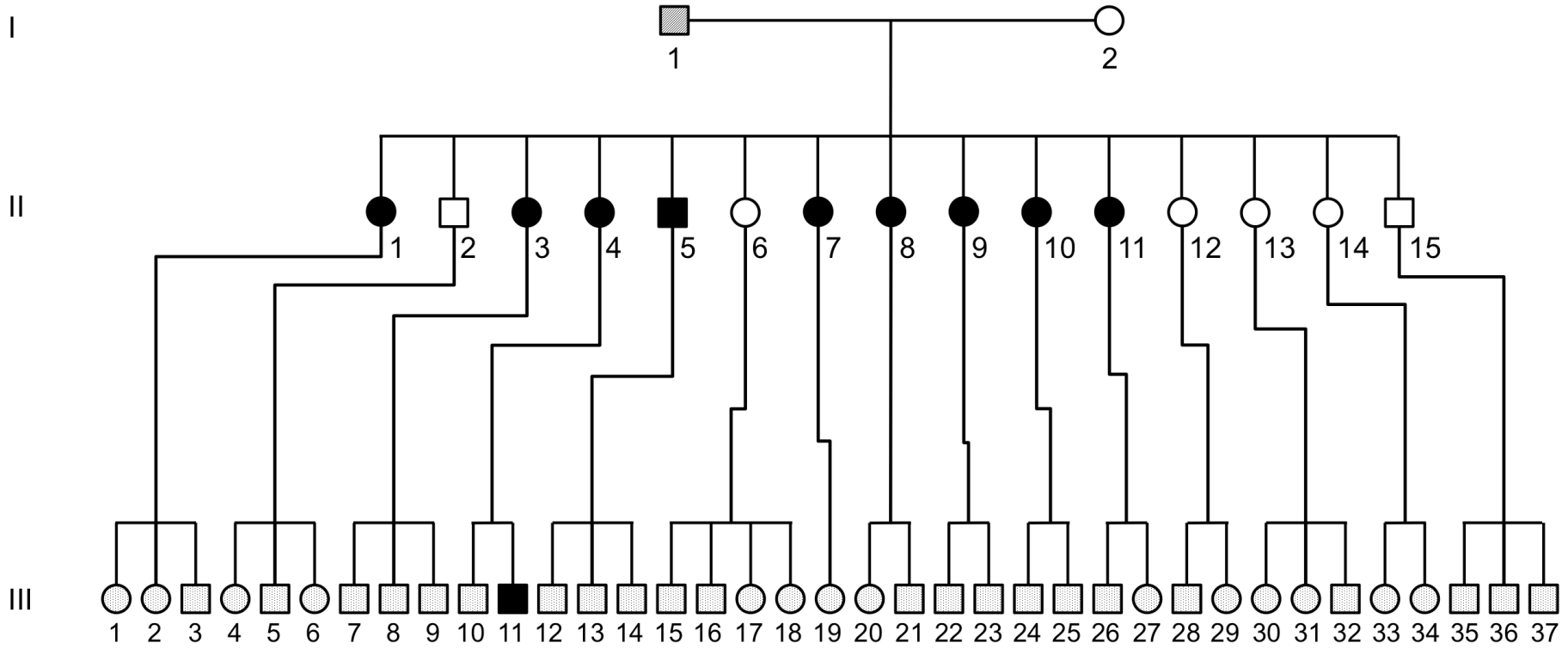

Figure 19 – Genetic pedigree of the index family. Affected individuals are shown by solid black symbols. Individual I:1 was reported to be affected but was now deceased and so this could not be confirmed. It should be noted that, although one individual in the third generation (individual III:11) showed signs of dystonia, the affectation of all other members of this generation was uncertain as they were all below the average age of onset of symptoms in the second generation. In the interests of space, the spouses of individuals in the second and third generations are not shown.

6.2.2 Exome Sequencing

Given that linkage would be uninformative in this family, we decided to perform exome sequencing on the DNA of three affected individuals in order to provide the best chances of being left with a workable number of variants. DNA from individuals II-5, II-8 and III-11 was thus used to perform whole exome sequencing as per section 4.14. Details of the filtering of exome sequencing data are given in the results section below.

6.2.3 Expression Data

Expression data for all of the variants discussed below were collated as per section 4.19.

6.3 Results

6.3.1 Exome Sequencing

Exome sequencing produced excellent breath and depth of coverage. Key coverage statistics – percentage of the target definition covered a 2x, 10x and 20x; mean read depth; and total number of variants detected – are summarized in table 15 for each sample.

Table 15 – Summary of exome coverage data from the three samples subjected to NGS.

Sample	Coverage x2	Coverage x10	Coverage x20	Mean Read Depth	Total Variants
1	89.2%	84.3%	78.3%	45	20838
2	92.7%	86.5%	81.4%	54	21266
3	90.6%	86.3%	80.3%	52	20907

6.3.2 Aggressive Filtration to Identify Candidate Causal Variants

The first step in the analysis of the data was to filter out all variants that were not shared between all the three affected individuals whose DNA had undergone next-generation sequencing. After sequential comparison and filtration, there remained 12,417 shared variants. Next, we removed any synonymous variants not affecting canonical spice sites, leaving 6,071 shared variants to carry over to the next stage. In accordance with our assumed model of autosomal dominance, we then removed all

homozygous variants, leaving 2,724 shared, heterozygous variants that might be expected to have a consequence on protein transcription or translation. In accordance with general conventions for variant filtration of exome data in kindreds exhibiting Mendelian diseases presumed to inherited in an autosomal dominant fashion, we then removed all variants that were present in the standard databases of natural human sequence variation (dbSNP, 1000 genomes, the NHLBI exome sequencing project, and complete genomics 69) at a minor allele frequency of greater than 0.1%. This left variants under consideration.

At this stage, it was clear that the number of variants left under consideration was too great. Thus, in order to reduce the list to a number that could be realistically tackled by manual curation, a more aggressive filtering strategy was required than would have been ideal had it been possible to narrow the genomic search space further by use of linkage data. We reasoned that, although cervical dystonia is not itself particularly rare, this particular form of cervical dystonia – highly-penetrant and dominantly-inherited – is only very rarely seen. We chose, therefore, to continue working on the data under the assumption that the variant underlying in this highly-penetrant form of dystonia would not have been recorded in databases of sequence variation to date. We recognize that this assumption is not unproblematic. This is especially so given the late-onset of symptoms in this kindred: carriers of a potentially causal mutation for such a condition could potentially have been included in projects designed to assess normal human genetic variation while asymptomatic, even though, at a later age, they might go on to manifest the disease. However, in order to be able to reduce the candidates under consideration to a manageable level, there seemed little choice but to filter out all variants that were not truly 'novel' – that is to say, to carry forward only variants that were not recorded in the aforementioned, commonly-used databases of natural human sequence variation. Finally, as an additional measure, we also removed any variants in regions of segmental duplication with greater than 95% sequence homology as experience has taught that these are almost always located in the non-transcribed pseudogene, which is naturally subject to greater genetic drift. After these operations, there remained 18 variants left for consideration.

6.3.3 Further Refinement of the List of Candidate Variants by Manual Curation

As obtaining primers for 18 exons, optimizing them and testing each of them for segregation in such a large family would be expensive, time consuming and, most importantly perhaps, a drain on patient DNA, an attempt was made to further refine the number of variants under consideration by gathering as much data as possible on each variant and the gene it was located in. This information was then used as the basis to assign each variant to one of two categories: either 'possible' candidate or 'unlikely' candidate. Any variant for which no data could be found to favor assignment to either category was left as a 'possible' candidate. It was understood that this process – being essentially qualitative – was necessarily imperfect (see discussion below) and would never be accepted as the basis of a publication on a new disease gene. However, at the same time, it offered a pragmatic way forward that essentially aimed to build into the decision-making process two characteristics of proposed disease-causing mutations that tend be associated with eventual confirmation of pathogenicity. Firstly, the proposed disease gene should be 'biologically plausible': 1) it should be expressed in the diseased tissue; 2) it should ideally have some link to the known molecular mechanism of that disease; and 3) it should not be associated with another disease, including recessive diseases where the heterozygous parents are reported to be neurologically normal; and 4) on review of the databases of normal human sequence variation, it should not ideally be littered with other variants that are predicted to be damaging, particularly if the putatively-associated disease trait is dominantly inherited, in which case the occurrence of annotated variants predicted to result in significant protein truncation or non-sense mediated decay would raise suspicion that the disease-gene association is spurious (unless, of course, a purely dominant negative mechanism of action was postulated). Secondly, the proposed variant should: 1) affect a base and/or amino acid that is evolutionarily conserved; 2) it should not universally be predicted to be benign or tolerated by multiple *in silico* prediction programs; and 3) it would, ideally, affect a known functional protein domain. Although it must be said upfront that, with the sole exception of an absence of gene expression in the diseased tissue, it is possible to find examples of a verified disease gene that breaks at least one of these rules, it remains the

case that most verified disease-causing genes and mutations do meet many if not all of these criteria.

Information curated and considered when making this decision included: 1) the Grantham score (a measure of the dissimilarity of two amino acids involved in a substitution); 2) *in silico* predictions of pathogenicity from three different programs that use diverse methods to generate their predictions; 3) three complementary measures of evolutionary conservation; 4) organism wide expression data; 5) data regarding protein function; 6) data regarding location of the variant within the protein with reference to known or predicted functional domaines, 7) previous associations with other diseases; and 8) the presence of other mutations likely to be deleterious in databases of normal human sequence variation (particularly stopgain mutations given the presumed mode of inheritence, as genes tolerating stop gains in normal individuals are unlikely to cause disease by haploinsufficiency). These data and the qualitative judgements made on the basis of it are summarized in table 16 and 17.

Using this data, 9 variants were designated as unlikely candidates and excluded from further analysis. Manual inspection of the raw exome sequencing data demonstated that 3 variants (in the genes *PKD1L2*, *HRC* and *ZNF677*) were actually polymorphisms, already annotated in dbSNP or 1000 Genomes that had not been identified in the initial, automated annotation process as the offending indel was incorrectly from the raw data (this is not an uncommon occurrence with indels in detected by NGS, especially if the affected area is repetitive, which makes accurate mapping to the reference genome considerably harder). The remaining rejected variants were either discarded as a result of one of a combination of the following factors (higlighed in red and emboldened in tables 16 and 17):

1) inherent lack of plausibility (*KRTAP1-3* and *HHLA2*)

2) negative conservation scores (*NELL2*, *FAM186A* and *MTPAP*)

3) universal predictions of toleration (*NELL2* and *FAM186A*)

4) definite association with another well characterized disease (*SH3TC2*)

5) genetic considerations related to similar mutations seen in data of normal sequence variation (*GALK2*).

Table 16 - Characterisation of candidate causal variants based on the nature of the genetic variant detected. Mutation code abbreviations are as follows: NS=nonsynonymous; FS=frameshift; NFS=nonframeshift; SG=stopgain; SNV=single nucleotide variant; DEL=deletion; INS=insertion. For *in silico* predictions of pathogenicity: D = deleterious (SIFT)/disease-causing (MutationTaster); T = tolerated; Prob = probably damaging; B = benign; Poly = polymorphism. NMD = predicted to lead to nonsense mediated decay.

Gene	Mutation Code	Mutation	Grantham score	SIFT	PolyPhen	MutationTaster	GERP Score	PhyloP Score	PhastCons Score
ZDHHC18	NS SNV	c.C461T:p.T154I		D - 0.02	Prob - 0.59	D - 0.999	5.27	3.584	1
CACNA2D3	NS SNV	c.A1814G:p.Y605C	194	D - 0.01	Prob - 0.947	D - 0.9999	6.16	6.99	1
FAM107A	NS SNV	c.G308A:p.R103Q	43	D - 0	Prob - 0.992	D - 0.99999	5.05	5.06	1
HHLA2	FS DEL		NMD						
NDUFA2	NS SNV	c.G50A:p.R17H	29	T - 0.06	B - 0.135	D - 0.9999999	4.93	2.27	1
ARAP3	NS SNV	c.G4001A:p.R1334Q	43	T - 0.11	Prob - 0.995	D - 0.999	5.66	7.24	1
SH3TC2	NS SNV	c.A1150T:p.I384F	21	D - 0.02	B - 0.258	D - 0.999	4.64	1.35	1
MTPAP	NS SNV	c.A1651G:p.I551V	29	**T - 0.3**	**B - 0.009**	**Poly - 0.9999**	3.49	4.44	0.973
NELL2	NS SNV	c.T83G:p.L28R	103		**B - 0.55**	**Poly - 0.9999**	**-2.38**	**0.08**	**0**
FAM186A	NS SNV	c.A4832T:p.Q1611L	113	**T - 0.71**	**B - 0**	**Poly - 0.9999**	**-7.33**	**-8.35**	**0**
SLC4A8	NS SNV	c.A1283G:p.N428S	46	T - 1	B - 0.007	D - 0.67	5.42	0.27	0.78
GALK2	SG SNV	c.975delC:p.Y325X	NMD						
PKD1L2	FS DEL	c.705_706del:p.235_236del	NMD						
NCOR1	NS SNV	c.T3331C:p.S1111P	74	D - 0.04	Prob - 0.937	D - 0.99999	6.07	7.13	1
KRTAP1-3	NFS INS	c.112_113insAGCTGCTGTGAGAC C:p.S38delinsSCCETS	Data not applicable to non-frameshift mutations						
HRC	NFS DEL	c.742_744del:p.248_248del							
ZNF677	FS INS	c.523_524insAT:p.K175fs	NMD						
EPB41L1	NS SNV	c.C1285T:p.R429C	180	D - 0	Prob - 0.976	D - 0.999999	5.48	3.76	1

Table 17 - Characterisation of variants based on gene function, biological plausibility, disease association and genetic factors.

Gene	Gene Function	Plausibility	Known Cause of Disease	Genetic
ZDHHC18	Palmitoyltransferase activity towards HRAS and LCK	Possible	No	
CACNA2D3	Subunit of voltage-dependent Ca2+ channels; regulates Ca2+ current density and activation/inactivation kinetics of channel.	High	No (link to pain)	
FAM107A	Suppresses cell growth when transfected into cells. May play a role in tumour development	Low	No (link to cancer)	
HHLA2	**Immunoglobulin**	Very low		
NDUFA2	Accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase	Low	No	
ARAP3	GTPase-activating protein that modulates actin cytoskeleton remodeling	Possible	No	
SH3TC2	Involved in Schwann cell maintenance	Low	**CMT4C (recessive), parent normal**	
MTPAP	Mitochondrial polymerase. Creates UAA stop codons that are not encoded in mtDNA	Low	No	
NELL2	Regulator of cell growth?	Low	No	
FAM186A	Unknown	Possible	No	
SLC4A8	Plays a major role in pH regulation in neurons.	Possible	No	
GALK2	Carbohydrate transport and metabolism	Low		**Multiple stop gains in dbSNP**
PKD1L2	May function as an ion-channel regulator. May function as a G-protein-coupled receptor.	Possible	No (link to behavioural disorders)	Annotated
NCOR1	Mediates transcriptional repression by certain nuclear receptors.	Possible	No (link to cancer)	
KRTAP1-3	**Essential for formation of rigid and resistant hair**	Very low	No	
HRC	Regulation of calcium sequestration or release in the SR of skeletal and cardiac muscle	Low	No	Annotated
ZNF677	Transcription regulation	Possible	No	Annotated
EPB41L1	May function to confer stability and plasticity to neuronal membrane.	High	No (linked to AD mental retardation)	

This left 8 variants for segregation analysis. Primers were thus designed to amplify the exon of each gene in which the detected variant was located and mutational status was ascertained for all individuals in generation II (where affectation could be assigned with reasonable certainty) for whom DNA was available. By definition, individuals II-9 and III-11 are affected. The results of this analysis are shown in the table 18.

Table 18 – Summary of segregation analysis of all members of the second generation for which DNA samples were available. The number of phenotype-genotype mismatches are shown in the far right column, assuming a fully-penetrant autosomal dominant inheritance pattern. M = mutation detected in the heterozygous state; wt = homozygous widetype alleles.

	II-3	II-4	II-5	II-7	II-8	II-10	II-11	II-12	II-13	II-14	Mismatches
Affected	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	
SLC4A8	M	M	M	M	M	M	wt	M	M	wt	3
FAM107A	wt	M	M	wt	wt	M	M	wt	wt	wt	2
NCOR1	wt	M	M	M	M	M	wt	wt	wt	M	3
ARAP3	wt	M	M	M	wt	M	wt	M	M	wt	4
ZDHHC18	wt	M	wt	wt	M	M	wt	wt	wt	wt	4
EPB4IL1	wt	M	wt	M	wt	M	wt	M	wt	wt	5
NDUFA2	wt	M	M	M	wt	M	wt	M	M	wt	5
CACNA2D3	M	M	M	M	wt	M	M	wt	wt	wt	1

As can be seen for the table above, no candidate variant segregated perfectly with disease in this family. The closest match was for *CACNA2D3*. This candidate gene had been ranked highly on the basis of its biological plausibility and on the characteristics of the variant detected within the gene.

6.3.4 An Overview of CACNA2D3

A limited amount of information exists on the gene *CACNA2D3*. This gene encodes a member of the α-2/δ subunit family, a protein involved in the voltage-dependent calcium channel complex. Voltage-gated calcium channels (VGCCs) are thought to exist in the plasma membrane as heteromeric proteins, in which the α1 subunit is associated with two auxiliary subunits, the intracellular β subunit and the extracellular $\alpha_2\delta$ subunit. These subunits influence the trafficking and properties of Ca_V1 and Ca_V2 channels. The first question therefore is whether the expression pattern of this gene

might be compatible with the development of a purely neurological disease. Even here things become complicated. According to Gong *et al.* (2001), *CACNA2D3* is only expressed in the brain, in the mouse at least [322]. This would be an extremely favourable expression pattern. According to Hanke *et al.* (2001), however, it is more widely expressed [323] and recent publications have suggested that methylation of *CACNA2D3* may be associated with a poorer prognosis in various cancers, such as that of the stomach, oesophagus, breast and brain [324-327]. Data from RNA sequencing experiments suggest a likely compromise: it is probably expressed at some level throughout the organism, but appears to preferentially expressed in the brain and heart. According to our own data (figure 20), it appears, within the brain, to be expressed at highest levels in the basal ganglia, followed closely by the cortex and hippocampus.

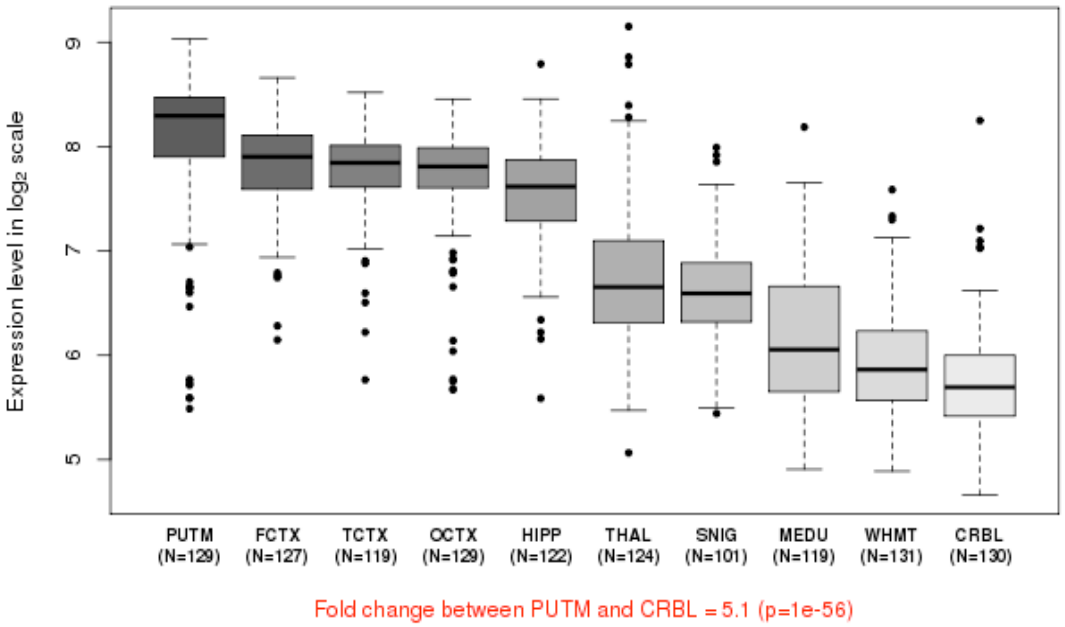

Figure 20 - Expression data for the gene *CACNA2D3* in man. The panel on the left shows the organism wide expression data based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, across the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

This expression pattern would be consistent with that of a gene capable of causing dystonia and is similar to the expression pattern of *ANO3*, which we found to be mutated in a family with autosomal dominant early onset tremulous cervical dystonia (covered in chapter 5), and also to that of *HPCA*, which we found to be mutated in a family with autosomal recessive generalised dystonia (covered in chapter 7)

The second question is whether the gene's function might fit in with currently accepted models of pathogeneisis for dystonia. Experimental research suggests that $\alpha_2\delta$ proteins modulate the activity of the VGCCs, which in turn drive neurotransmitter release in neurons. Specifically, α2δ subunits set synaptic abundance of VGCCs and $\alpha_2\delta$s configure synaptic VGCCs to drive exocytosis through an extracellular metal ion-dependent adhesion site (MIDAS), a conserved set of amino acids within the $\alpha_2\delta$ protein's predicted von Willebrand A (VWA) domain [328]. Expression of $\alpha_2\delta$ with an intact MIDAS motif leads to an 80% increase in release probability, while simultaneously protecting exocytosis from blockade by an intracellular Ca^{2+} chelator [328]. Thus, mutation of this gene could potentially alter neuronal excitability or propensity to neurotransmitter release in an areas of the brain associated with movement disorders.

The third question is whether the mutation we detected is located within an area of the protein known to have functional importance. The $\alpha_2\delta$ subunits have been described as type I transmembrane proteins, because they have an N-terminal signal peptide and a C-terminal hydrophobic and potentially transmembrane region. However, evidence has accumulated recently to suggest that $\alpha_2\delta$ proteins are probably associated with the plasma membrane through a glycosylphosphatidylinositol (GPI) anchor attached to δ part of the protein rather than the transmembrane domain [329]. The process of anchoring occurs within the endoplasmic reticulum, and it involves cleavage of the C-terminal hydrophobic peptide at the ω-residue and attachment of a GPI group to this residue, which then attaches the protein to the membrane through its lipid side chains. The hypothesis was originally prompted by the results of prediction programs examining the $\alpha_2\delta_3$ proteins alone, but its validity is now supported by good experimental evidence for all $\alpha_2\delta$ proteins [329]. The mutation we detected occurs in the

α_2 domain of the protein and so it unlikely to affect GPI-mediated membrane binding. The other key domain within the protein is the MIDAS domain (mentioned above), which binds divalent metal cations and its function is required for $\alpha_2\delta$ proteins to effectively modulate calcium currents. The mutation that we detected is located approximately 400 amino acids upstream from this motif and, thus, altered binding of divalent cations cannot be invoked as a likely causal mechanism. Overall, the mutation that we detected does not appear to disrupt area of the protein with previously described critical role in its function. This, of course, does not preclude the idea that it may affect an area with as-of-yet unknown functional significance.

The final question revolves around whether the gene has been associated with any particular function within the nervous system already. There are only two papers that suggest a possible functional role for *CACNA2D3* within the nervous system. The first links the gene to nociceptive processing. Using genome-wide neuronal-specific RNAi knock-down in Drosophila, Neely *et al.* (2010) set about identifying novel genes implicated in heat nociception in the fly [330]. One of the genes that they identified by this means was the calcium channel subunit *straightjacket*. The mammalian ortholog of *straightjacket* is *CACNA2D3* and its role in thermal nociception was confirmed in knock-out mice, which exhibit a significantly impaired basal heat pain sensitivity and delayed thermal hyperalgesia after inflammation [330]. They went on to pinpoint single nucleotide polymorphisms in human *CACNA2D3* that are associated with reduced acute heat pain sensitivity in healthy volunteers, as well as chronic postsurgical back pain [330]. This finding is intriguing since research conducted after the original publication of *ANO3* as a dystonia gene has suggested that it too may be involved in nocioceptive processing. Within the dorsal root ganglia, *ANO3* is expressed mainly in the IB4 positive, non-peptidergic nociceptors, which also co-express the sodium-activated potassium (KNa) channel Slack [331]. In rats, the expression of *ano3* appears to promote KNa channel activity and thus dampen neuronal excitability. Dorsal root ganglia neurons from *ano3* knock-out rats have reduced Slack expression, broadened action potentials and increased excitability [331]. Moreover, the *ano3* knock-out rats exhibit increased thermal and mechanical sensitivity [331]. The data thus suggests that both human *ANO3* and *CACNA2D3* may be involved in the processing of pain signals

and the two may therefore be envisaged as sharing a common function as modulators of neuronal excitability. Continuing with the theme of modulation of signalling pathways, the second paper reporting on *CACNA2D3* function suggests that the gene may also be important in modulating neuronal signals within the auditory pathway. *CACNA2D3* knock-out mice display a reduced auditory startle response, most likely as a result of impaired synaptic transmission. The authors further demonstrated that lack of *CACNA2D3* resulted, on a cellular level, in a reduction of $Ca_v2.1$ expression in the somata of spiral ganglion neurons, but also in their synaptic terminals contacting bushy cells. Given this data, it does not seem too implausible to speculate that mutations in *CACNA2D3* might be envisioned to cause dystonia by 'altering the gain' in neuronal pathways controlling movement within the basal ganglia (where it is highly expressed) in a manner similar to that we hypothesised to underlie the effect of *ANO3* mutations.

6.4 Discussion

We set out to identify the genetic cause of autosomal dominant, late-onset tremulous dystonia in an unusually-large, English kindred. Despite the attractions offered by the unusually large size of the second generation and the seemingly near-complete age-related penetrance of the disorder, the family was always going to be a challenge to solve. The main reason for this is that the late-onset of the disorder means that the affectation status of the third generation could not be accurately be determined. By extension, no linkage analysis is possible and thus the genomic search space is unmanageable.

The first concession made to this imperfect situation was to only consider novel variants. Although any discoveries based on such a concession would always have been unpublishable without further supporting evidence (in the form of several independent segregating kindreds), we felt that it was, nonetheless, worth pursuing this avenue in order to determine if an obvious candidate variant stood out. As it happened, no candidate variant either immediately stood out as perfect or indeed survived a later segregation analysis. All that can be said is that, of the candidate-causal, novel variants that had been selected for segregation analysis, that in which the pattern of inheritance matched most closely the pattern of disease in the family was also the same variant that

was deemed to have a higher *a priori* probability of being causal on the basis of the (necessarily somewhat subjective) character profile compiled for it – that is to say, the p.Y605C variant in the gene *CACNA2D3*. However, since, even in the case of this variant, segregation analysis still showed one mismatch (and notwithstanding the fact that the gene is extremely large), a decision was made not to go on to sequence the gene in an independent dystonia cohort. At the same, this family does show enormous potential from a genetic point of view. If you will allow me the liberty of an everyday analogy: like a good wine, it will mature with time. In about 20 years or so, when the affectation status of the 3rd generation can be more easily ascertained, finding the gene should be relatively easy, given the number of the affected and unaffected individuals in the second and by then third generation.

Could mutations in *CACNA2D3* be the cause of dystonia in this family? It is possible. The expression pattern is favourable and the genes function shows similarities to that of *ANO3*. If one were to accept this hypothesis, however, then one would have to simultaneously accept the hypothesis that individual II-8 is a phenocopy. There are three main reasons why such an individual may well turn out to be a phenocopy and both are particularly applicable to dystonia.

The first reason revolves around the fact that some diseases are not so uncommon in the general population that it is always possible that an individual might suffer from that disease, but that it might not be the result of a genetic mutation that is causal in the rest of the family. This exactly situation occurred during the analysis of the kindred in which the Parkinons' gene *SNCA* was initially identified. For reasons mentioned later, prevalence estimates for cervical dystonia are wide ranging. The most robust studies suggest a prevalence of about 28 – 183 cases per million of population [332]. Whilst unlikely, it is all the same within the realms of possibility that individual II-8 is affected by cervical dystonia, but not as a result of the mutation carried by the rest of her affected siblings.

The second reason is linked to the difficulty in being certain that a clinical diagnosis is accurate in the case of certain disorders. For disorders in which the clinical picture is

so strikingly obvious or, alternatively, where a diagnostic test exists that can accurately separate affected from unaffected individuals (for example, a DaT scan in Parkinson's disease), there is usually little difficulty is determining whether an individual's symptoms are due to the disorder being examined. Dystonia, however, is a particularly difficult condition to deal with in this respect. Not only does no robust diagnostic test exist, which might determine whether an individual is truly affected or not, but the clinical phenomenology of the condition is extremely variable, ranging from disabling, deforming and grossly abnormal movements of the head, trunk and limbs to subtle, relatively unintrusive, and easily overlooked movements or postures of a single or limited part of the body. Enlisting movement disorders specialists – as we did here – to phenotype the family can help, but such individuals – highly attuned as they are to the tiniest suspicious movement – probably serve best to increase the positive predictive value of a clinical diagnosis, rather than reduce its false positive rate. Nonetheless, it would only be fair to say that I have reviewed the examination video of individual II-8 several times and clinical diagnosis of cervical dystonia was certainly not made on the basis of the subtle signs.

The third and final reason for phenocopies is that an apparently affected individual's clinical signs are a learned behaviour with a psychological basis. On this vexed topic, I will go no further than to say three things: 1) that functional illness is a sufficiently common diagnosis that it probably enters most practicing neurologist's heads at least once a week; 2) that movement disorders are particularly common and difficult manifestations of functional disease, as decisively diagnostic tests are often lacking; and 3) that I was warned by one of my supervisor's, on surveying the family tree, that the kindred looked 'over-dominant' (his words, not mine) and that he felt it was likely that there were phenocopies amongst the supposedly affected individuals.

With all that said, given the concessions that I was forced to make in selecting my candidate variants, it is equally possible that: 1) I have filtered out a low frequency, but nonetheless causal variant; 2) I have discounted a novel, but seemingly implausible variant and therefore not subjected it to segregation analysis; 3) the causal variant is novel and exonic, but has not been covered in one or both exomes; or 4) the causal

variant is not exonic or is a large scale deletion/duplication or exonic copy number variant, all of which are impossible/difficult to detect by means of exome sequencing.

In summary, therefore, work on this kindred must, ultimately, be declared a failure. Its inclusion in this thesis, however, is – in my opinion at least – important, as it illustrates some of the key reasons for which exome sequencing studies often grind to a halt, a subject to which I will return in my concluding remarks. Perhaps understandably, the majority of this thesis is taken up with describing my relative successes, but it is, at the same time, important to be honest and thus make the following confession: for every family that I present herein, in which I either identify a by-now-already-published or an unpublished-but-altogether-plausible causal mutation, you can be assured that there was almost certainly one or more families in which work ground to a halt, either at this or at an even less advanced stage.

CHAPTER 7:

Exome Sequencing in Autosomal Recessive Generalised Dystonia

7. Exome Sequencing in Autosomal Recessive Generalised Dystonia

7.1 Introduction

At the time of completion of this work, recent rapid advances in the field of dystonia genetics had pushed the total number of known causative genes for primary isolated dystonia to six (*TOR1A* [136], *THAP1* [157], *CIZ1* [181], *ANO3* [333] , *GNAL* [334] and *TUBB4A* [186, 187]). In all cases, the dystonia resulting from mutations in these genes is transmitted in an autosomal dominant manner, often with reduced penetrance. However, within the parallel DYT loci classification system, there remained somewhat of an anomaly: DYT2. Unlike every other dystonia locus, DYT2 was never defined on the basis of a linkage interval, but merely on the observation of a phenotype: that is, primary, isolated dystonia that appeared – unusually with respect to this condition – to be inherited in an autosomal recessive fashion [335]. Perhaps regrettably, as further putatively-recessive, isolated dystonia kindreds were reported in the literature, there was a tendency to simply lump them together on the basis of their presumed mode of inheritance under the banner of DYT2 or 'DYT2-like' dystonia [124, 336, 337], even though there was no evidence to suggest that they all shared the same genetic aetiology. The sole exception to this practice was the creation in 2008 of the locus DYT17 to designate a region on chromosome 20, defined by homozygosity mapping, in a consanguineous Lebanese kindred exhibiting apparently autosomal-recessive, isolated dystonia [338]. Yet, as with DYT2, the causal gene remains to be identified [338].

The genetics of DYT2 and, more generally, autosomal recessive, isolated dystonia has thus remained a persistent mystery in the field of dystonia research. In part at least, this is due to the fact that this form of dystonia appears to be relatively rare, as evidenced by the very limited reports in the literature [124, 335-338]. Some have even expressed doubt as to whether an autosomal recessively inherited form of isolated dystonia exists at all [123, 339]. In reality, if autosomal recessive isolated dystonia were proven to exist, it would be expected that it would show the greatest prevalence in populations characterised by a higher frequency of consanguineous marriages. Elsewhere, it would be expected to be relatively rare. Even when it did arise, it might not be recognised as a recessive disorder, since the tendency towards small family size in

much of the Western world means that singleton cases could easily be mislabelled as a sporadic or assumed to be the result of cryptic inheritance of or *de novo* mutation in one of the known autosomal dominant dystonia genes (especially in view of the markedly reduced penetrance and wide clinical expressivity of the most common of these, *TOR1A*). Thus, globally, autosomal recessive isolated dystonia may well be less rare than the scientific literature to date suggests. In any case, identifying a genetic cause for 'DYT2 dystonia' remains important, as it would: 1) firmly establish the existence of an autosomal recessive form of primary isolated dystonia; 2) allow the anomalous DYT2 locus to be linked to a causative gene; 3) permit further genetic screening to determine whether other kindreds published under the banner of DYT2 dystonia actually share a common genetic cause; and, most importantly of all, 4) further expand our knowledge of the molecular pathways involved in the pathogenesis of dystonia.

With the advent of next generation sequencing technologies, we revisited a Sephardic Jewish kindred, previously reported as 'DYT2-like', exhibiting autosomal recessive, isolated dystonia [124]. The three affected siblings (now aged 61, 57 and 51, respectively) were the product of a consanguineous marriage between two first cousins. Both parents were neurologically normal and there was no report of any dystonia within the wider kindred. A combination of autozygoisty mapping and whole exome sequencing were employed to identify the likely causal variant in this family.

7.2 Subjects, Materials and Methods

7.2.1 Clinical Details of the Index Family

The index family has previously been described extensively elsewhere and readers are referred to that description for full details of the clinical presentation and previous investigation [124]. A brief *précis* of this information is provided below for the purpose of this thesis. Figure 21 shows the genetic pedigree for the core family. With the exception of the affected siblings, there was no report of dystonia within the extended family in the previous two generations (the extended family tree has previously been published in the paper by Khan *et al.* [2003]) [124].

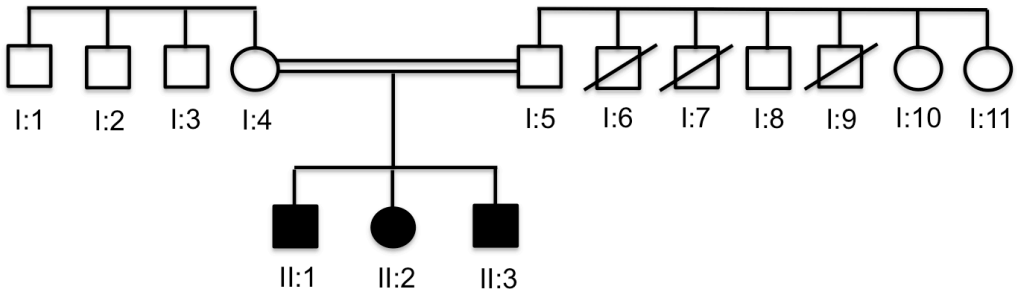

Figure 21 - Genetic pedigree of the index family showing a consanguineous marriage resuting in three siblings affected with young-onset, generalised dystonia. Both parents and all members of the wider family were free of neurological disease.

This Sephardic Jewish family originated from Tehran in Iran. Intermarriage was relatively common in their small community and the unaffected parents (I:4 and I:5) are first cousins. Their three children (II:1 through II:3) were all affected, consistent with autosomal recessive inheritance. All affected individuals had normal birth and developmental milestones. There was no history of drug use or psychiatric symptoms and no diurnal variation or worsening with intercurrent illness. None had cerebellar, pyramidal, or extrapyramidal signs.

Individual II:1 – Initial signs of a dystonic disorder were noted around one year after birth, when the patient was noted to walk 'pidgeon-toed'. At age 7 she reported intermittent jerking of the head and unclear speech. Her symptoms progressed over the next 6 years, with the development of torticollis, facial grimacing, blepharospasm, intermittent dysphagia, and breathlessness. At age 40 years, she developed mandibular muscle spasms with teeth grinding. On clinical examination, there was marked dystonic posturing and jerking of the upper limbs with blepharospasm and leftward torticollis.

Individual II:2 – Initial symptoms consistent with dystonia where noted at the age of 8 years of age, when the parents observed abnormal postures, including intermittent

inturning of the left foot and dragging of the left leg. At age12, he reported abnormal postures of the arms, difficulty writing, and an intermittent intention tremor. In his teenage years, he developed torticollis and dysarthria with no significant progression since then. On clinical examination, there was evidence of an exaggerated lumbar lordosis, a dystonic gait with in-turning of the feet. There was dystonic posturing of hands with choreoathetoid movements of the right arm and retrocollis. Speech was dysarthric.

Individual II:3 – Initial symptoms were observed in this woman at the age of 5, when she exhibited mild gait abnormalities compatible with dystonia. At 13 she developed painful jerky torticollis, intermittent shoulder movements, and tremor of both hands. At 16 she had marked grimacing, blepharospasm and lingual dystonia. On clinical examination, speech was dysphonic with laryngeal adductor spasms and respiratory dyskinesia. There was a "yes-yes" head tremor, right torticollis with retrocollic jerks, involuntary movements around the eyes and mouth, blepharospasm, and facial grimacing. There was marked dystonic posturing of the limbs.

In summary, all three affected siblings developed dystonia in their first decade of life, which gradually generalised over time, but remained most marked in upper limbs, cervical, and cranial regions.

Initially, the siblings were reported to be suffering from an atypical form of metachromatic leukodystrophy (MLD) on the basis of markedly reduced levels of arylsulfatase A in fibroblasts and leucocytes, reduced nerve conduction velocities and detection of brown metachromatic granules in sural nerve biopsies. Subsequent mutational screening by Sanger sequencing demonstrated that both the mother and the three siblings were homozygous and the father heterozygous for two variants in a *cis* configuration (NM000487.5: c.[1055A>G; *96A>G]) in the gene *ARSA* (MIM 607574), which are commonly referred to collectively as the 'polyA mutation'. The polyA mutation results in reduced levels of arylsulfatase A on biochemical assay without clinical symptoms (a state termed pseudodeficiency). No other mutations were detected in the remainder of the gene at that time or in the current study. In this context, the

detection of metachromatic granules in the sural nerve biopsy is unusual. Nonetheless, despite prolonged follow up, no clinical or radiological features of progressive central or peripheral demyelination have developed, making MLD highly unlikely. On current examination, there are no other neurological features besides the dystonia detectable on clinical examination and exhaustive radiological and biochemical investigations had failed to reveal any underlying cause. Extensive genetic testing – including but not confined to *TOR1A*, *THAP1*, *GNAL* and *ANO3* – has not revealed a causal mutation in these genes. Although the severity of the dystonia had gradually increased over time, the clinical course appears relatively benign: all three of the affected siblings continue to function well in daily life; there is no significant limitation of ambulation; and fixed deformities have not developed.

7.2.2 Selection of Cohorts for Mutational Screening

Subsequent dystonia cases used to search for further confirmatory mutations in candidate genes were drawn from a bank DNA samples donated with research consent, which are held at our institution. Despite the extensive nature of this clinical resource, the rarity of autosomal recessive, isolated dystonia meant that there were no other samples available from any other dystonia kindred where the inheritance pattern could definitively be said to be autosomal recessive. We were therefore forced instead to select cases where the history was merely 'not incompatible' with autosomal recessive inheritance (i.e. with either no family history or a family history of affected siblings only). Thus, for the initial cohort of 150 cases used to screen for mutations in the relevant exons of both candidate genes, we selected samples on the basis of an associated clinical history of the onset of isolated dystonia at 30 years of age or younger with no family history that might be taken to suggest autosomal dominant transmission, even with reduced penetrance. We made a particular effort to include all research samples where the associated clinical details made mention of a possible family history involving only siblings, but in practice this only amounted to a handful of samples and most cases were thus singletons. Because of the preponderance of singletons, it is, of course, inevitable that a proportion of the cases that we selected will have been either non-genetic or the result of cryptic autosomal dominant inheritance,

secondary to issues such as reduced penetrance, *de novo* mutations and germline mocaisism.

For the second cohort of an additional 288 cases, used in the mutational screening of all coding exons of *HPCA*, we employed a slightly different approach. As previously, we first identified those samples with an associated clinical description suggesting primary, isolated dystonia and no family history suggestive of autosomal dominant inheritance. From this pool, we then selected samples for actual sequencing on the basis of an associated clinical description of dystonia with a distribution similar to the that seen in the two families with recessive mutations in *HPCA* that we had by that time identified. That is to say, those with predominant upper limb, cervical or cranial involvement or those in whom the dystonia had generalised. Regrettably, the clinical details associated with many of these samples did not always include a clear indication of the age at onset. However, we did make an effort to preferentially include samples from individuals with a younger age of onset (as might be expected in an autosomal recessive disorder) by selecting only those samples where the date of banking was within 5 decades of the sample donor's date of birth (i.e. individuals aged 40 years or younger when the sample was donated).

All research samples used in the study had previously undergone diagnostic testing for mutations in *TOR1A* and were negative. As most of the cases had no family history, systematic testing for *THAP1* mutations had not been carried out. Finally, we also attempted to contact the authors of previous reports of autosomal recessive isolated dystonia in the literature in order to see if we could obtain DNA samples from affected members of these kindreds to include in our screening efforts. In the end, however, we were only able to obtain DNA from an affected member of the DYT17 kindred, which was screened for mutations in both exon 7 of *LAPTM5* and the whole coding region of *HPCA*, despite the fact that homozygosity mapping in this family had already suggested that the genetic cause was located elsewhere.

The study was approved by the relevant local ethics committee at our institution and informed consent was provided by all participants in accordance with its guidelines.

7.2.3 Whole Exome Sequencing and Generation of Coverage Statistics

DNA was extracted from whole blood samples obtained from all three affected siblings and both parents. DNA from one affected sibling was used to perform whole exome sequencing using Illumina's TruSeq (62Mb) DNA sample preparation and exome enrichment kits as per section 4.15. Reads were subsequently aligned and annotated using the standard approaches set out in sections 4.16.

Coverage statistics for the exome as a whole were obtained using Picard. Coverage across regions of shared homozygosity was calculated using BedTools against the CCDS definition of the exome and expressed as the percentage of bases target covered by at least one read.

7.2.4 Genotyping and Subsequent Homozygosity Mapping

Genome-wide genotyping data was obtained for all three affected siblings using the OmniExpress platform, which utilises approximately 500,000 markers, as per section 4.12. Autozygosity mapping was performed as per section 4.13.

7.2.5 Filtration of Variants Detected by Exome Sequencing

In view of the apparently recessive inheritance pattern and history of consanguinity, we initially selected all homozygous variants for consideration. Subsequently, synonymous variants not likely to affect a canonical splice site (i.e. those not within 10 bases, in either direction, of the intron/exon boundary) were discarded. Given the rarity of autosomal recessive dystonia, we further hypothesised that the causal variant was unlikely to be found in any database of normal sequence variation. However, in order to minimize the possibility of incorrectly assigning causality, we filtered out those variants that were recorded in the 1000 Genomes, NHLBI or Complete Genomics 69 datasets at a minor allele frequency of greater than 0.5%. Given that a variant found at even this frequency would be expected to be occur naturally in the homozygous state in around 1 in every 160,000 births, it seemed distinctly unlikely that we would risk filtering out the causal variant with this cut-off. Subsequently, variants that were located in a region of shared homozygosity were selected as potentially causal. No

filtration was performed on the basis of *in silico* predictions of pathogenicity or conservation scores.

7.2.6 Confirmation of Potentially Causal Variants and Subsequent Sequencing of Candidate in Independent Dystonia Cohorts

Primers were designed to amplify the relevant exons and exon/intron boundaries of both genes containing potentially causal variants, as defined above. For both variants, mutational analysis by chain termination methodology was first performed in the remaining siblings and both parents in order to verify an appropriate pattern of inheritance. Subsequently, the same primers were used to sequence the relevant exon of both genes in a cohort of 150 individuals with young-onset dystonia, selected as described in section 7.2.2. After a second potentially causal mutation in *HPCA* was detected in one of these samples, primers were designed to amplify and sequence the remaining coding exons of that gene in these 150 samples and a further cohort of 288 individuals with dystonia, selected as described in section 7.2.2. Exon 1, which is entirely untranslated, was not sequenced.

7.2.7 Generation of Nucleotide Multispecies Protein Alignments

In order to assess conservation of amino acid sequence, Uniprot, FlyBase and WormBase were interrogated for orthologous protein sequences. Actual alignment of the sequences obtained in this manner was performed using the freely available web-based application ClustalΩ.

7.2.8 Generation of Regional Gene Expression Data

Information on organism wide and brain region specific gene expression was collated as per section 4.19.

7.2.9 Generation of Rat-Neuron Primary Culture and shRNA Knockdown Rat Primary Cortical Cultures

Animal husbandry and experimental procedures were performed in full compliance with the United Kingdom Animal (Scientific Procedures) Act of 1986. For primary cortical cultures, *Wistar* wt rat pups were culled between postnatal days P1–3 and a

primary co-culture was prepared as described elsewhere [340]. Cerebral hemispheres were trypsinized and resuspended in 2 ml warm complete Neurobasal A medium and the cell suspension was plated on poly-L-lysine coated coverslips. The cultures were incubated in a humidified incubator at 37°C with 5% CO_2 in air for 3–4 h, then 2 ml pre-warmed complete Neurobasal A medium was added. All live cell imaging experiments were performed between DIV 10-14.

7.2.10 Rat hpca Knock-Down

The *hpca* knock-down in rat cortical primary cultures was performed by Effectene transfection (Qiagen) after 9-10 days in culture with either one or a pool of 4 specific-shRNAs targeting rat *hpca* (Thermo Fischer). These four individual shRNAs as well as a pool thereof were used to knock-down *hpca* in rat cells. The empty vector (pGIPZ) and the vector expressing a non-targeting RNA (SCR) were used as controls. Knock-down was confirmed by Western bot (please see results section for caveat regarding this and interpretation of results). The transfection was made following the manufacturer instructions and the sequences from the mouse *hpca* shRNA were as shown in table 19. 48 hours after transfection the cells were ready for subsequent experiments.

Table 19 – shRNAs used for *hpca* knock-down are shown with corresponding genetic sequences

shRNA label	Genetic Sequence
shRNA #1	AAAGAATACAGAACCTGAC
shRNA #2	TTTCTAGCATCTCCTCCCG
shRNA #3	CTTCTTGAACTCATCCACG
shRNA #4	TGTTTGTGTCCATTTGGCG

7.2.11 Functional Studies in Rat-Neuron Primary Culture

Astrocytes and neurons from primary cortical co-culture were loaded for 30 min at room temperature with 5 μM fura-2 AM and 0.005% pluronic acid in a HEPES-buffered salt solution composed of 156 mM NaCl, 3 mM KCl, 2 mM $MgSO_4$, 1.25 mM KH_2PO_4, 2 mM $CaCl_2$, 10 mM glucose, and 10 mM HEPES (pH was adjusted to 7.35 with NaOH). Fluorescence measurements were obtained on an epifluorescence inverted microscope equipped with a 20× fluorite objective. $[Ca^{2+}]_c$ was monitored in

single cells with excitation light provided by a Xenon arc lamp, and the beam passed through a monochromator at 340 and 380 nm (Cairn Research, Kent, UK). Emitted fluorescence light was reflected through a 515 nm longpass filter to a charge-coupled-device camera (Retiga, QImaging, Surrey, BC, Canada) and digitized to a 12 bit resolution. All imaging data were collected and analysed with software from Andor IQ (Belfast, UK). The fura-2 data were not calibrated in terms of $[Ca^{2+}]_c$ because of the uncertainty arising from the use of different calibration techniques. Areas for the analysis were chosen depending on the GFP fluorescence intensity and four independent experiments were performed for each condition. The number of cells analysed for each set of experiments is indicated in the result section and in the legends of the figures.

7.3 Results

7.3.1 Exome Sequencing and Homozygosity Mapping

Exome sequencing produced good coverage of the target. Using the TruSeq exome definition as a reference, coverage was 96% at a read depth of 2, 87% at a read depth of 10 and 74% at a read depth of 20. Mean read depth across the exome was 58. In total, 22,097 variants were detected.

The runs of homozygosity shared between all affected siblings are summarized in the Table 20, along with the number of Consensus Coding Database of Sequences (CCDS) genes in the region (obtained via Ensemble's Biomart), their percentage coverage and the number of variants detected in that region prior to any filtering. Overall, coverage of homozygous regions was well within the range expected for exome sequencing (≈85% of targeted bases). Moreover, manual inspection of the uncovered bases within these regions revealed that the vast majority were in the untranslated regions of genes, were exome sequencing tends to perform more poorly. Coverage of coding bases alone, where pathogenic mutations are more likely to be found, was thus better than indicated.

Table 20 - Regions of homozygoisty shared between all three siblings with genomic coordinates (hg19), length in megabases (Mb), the number of CCDS genes that lie in the region, the percentage of CCDS bases (including untranslated regions) covered by exome sequencing, the number of variants detected in that region and, finally, the number of potentially causal variants that remained after filtration as detailed in the methods.

Chr	Start	End	Length (Mb)	CCDS Genes	% Coverage	Variants Detected	Potentially Causal
1	12880356	20476391	7.60	86	84.3%	166	0
1	26909765	34686130	7.78	102	91.0%	43	2
3	126380804	127502549	1.12	7	97.9%	5	0
5	98552184	99968045	1.42	1	100%	0	0
6	34502022	36226525	1.72	28	95.9%	16	0
7	64926823	66464764	1.54	10	82.7%	3	0
8	48639976	49656604	1.02	5	80.5%	3	0
8	85802488	86990451	1.19	8	89.7%	4	0
11	47976882	51591253	3.61	13	95.9%	32	0
11	54794237	55943322	1.15	25	89.4%	52	0

7.3.2 Filtration of the Data and the Identification of Two Potentially Causal Variants

Filtering of the variants, carried out as described in the section 7.2.5, left just 2 possible candidate causal variants for further consideration (Table 20 and 21). Both were located in the largest stretch of shared homozygosity on chromosome 1. The first was missense mutation (c.625G>A, p.V209M) in exon 7 of *LAPTM5* (NM_006762), which encodes a lysosomal transmembrane protein. The second, also a missense mutation (c.225C>A; p.N75K), was located in exon 2 of *HPCA* (NM_002143), a gene which encodes the neuronal calcium-sensor protein known as hippocalcin.

According to Uniprot, the p.N75K variant in *HPCA* is located in the second EF-hand domain of the protein and affects an amino acid that is known to be critical to the domain's calcium binding properties. The variant detected in *LAPTM5* does not lie in

any annotated functional domain of the protein, according to Uniprot. Nonetheless, both variants are predicted to be highly damaging by all four in-silico prediction programs (SIFT, Provean, PolyPhen2 and MutationTaster) and the PhyloP and PhastCons scores suggest that both mutations affect a conserved base (see table 21 for scores). With regard to the affected amino acids, multiple species alignment of orthologous protein sequences (see figure 22) demonstrated that the amino acid affected by the p.V209M mutation in *LAPTM5* is conserved in most species for which an ortholog exists, with the exception of the tropical frog (*X. tropicalis*) which has an isoleucine at this position. However, no orthologs are known for worm (*C. elegans*), fly (*D. meganolaster*), zebra fish (*D. rerio*) or pig (*S. scrofula*). As might be expected given the affected amino acid's critical role in calcium binding, multispecies alignment for the protein sequence surrounding the p.N75K mutation in *HPCA* showed absolute interspecies conservation of the affected amino acid down to fly and zebra fish. Again, no orthologue is known to exist in the genome of *C. elegans*.

7.3.3 Expression Data Supporting HPCA as a Higher-Priority Candidate for Dystonia

Publically-available expression data based on experiments using EST tags demonstrated that both *LAPTM5* and *HPCA* are expressed in a highly tissue-specific manner, though with quite distinct patterns of expression (Fig 23A and 23B). Expression of *LAPTM5* was detected predominantly in the cells of the blood and associated haemopoietic organs. This data is in line with the published murine tissue expression analysis performed at the time of the gene's discovery, which showed it to be present at the highest levels in haematopoietic tissues (peripheral blood leucocytes, thymus and spleen) and the lung (possibly from contamination with alveolar macrophages), present in low levels in the placenta, liver and kidney, and absent or barely detectable in all

Table 21 – Summary of Candidate Causal Variants after Filtration

Summary of both candidate causal variants with conservation scores (PhyloP and PhastCons), *in silico* predictions of pathogenicity (SIFT, Provean, PolyPhen2 and MutationTaster) and the location of the variant with respect to predicted functional domains of the protein (from Uniprot). Previously reported refers to the whether the variant can be found in the databases of dbSNP, NHLBI Exome Sequencing Project, the 1000 Genomes Project, and Complete Genomics 69. Both variants are conserved and predicted to be damaging by all four prediction programs. Actual numerical scores provided by the *in silico* prediction programs are shown here in brackets for the sake of completeness and readers are referred to the programs websites for a detailed explanation of their meaning. D = damaging; C = conserved

Chr	Position (hg19)	Gene (Transcript)	Homozygous Change	Previously reported	PhyloP	Phast Cons	SIFT	Provean	PolyPhen	Mutation Taster	Functional domain of protein
1	31208094	*LAPTM5* (NM_006762.2)	c.625G>A p.Val209Met	No	C (3.37)	C (1)	D (0.001)	D (-2.7)	D (0.999)	D (0.986)	No
1	33354724	*HPCA* (NM_002143.2)	c.225C>A p.Asn75Lys	No	C (2.71)	C (1)	D (0.001)	D (-5.239)	D (0.993)	D (0.999)	Yes: EF-hand domain 2

other tissues (including, most importantly, the brain) [341]. *HPCA*, on the other hand, appears to be expressed almost exclusively in the brain. This would be more in keeping with the expression pattern of a gene capable of causing a purely neurological disorder like isolated dystonia.

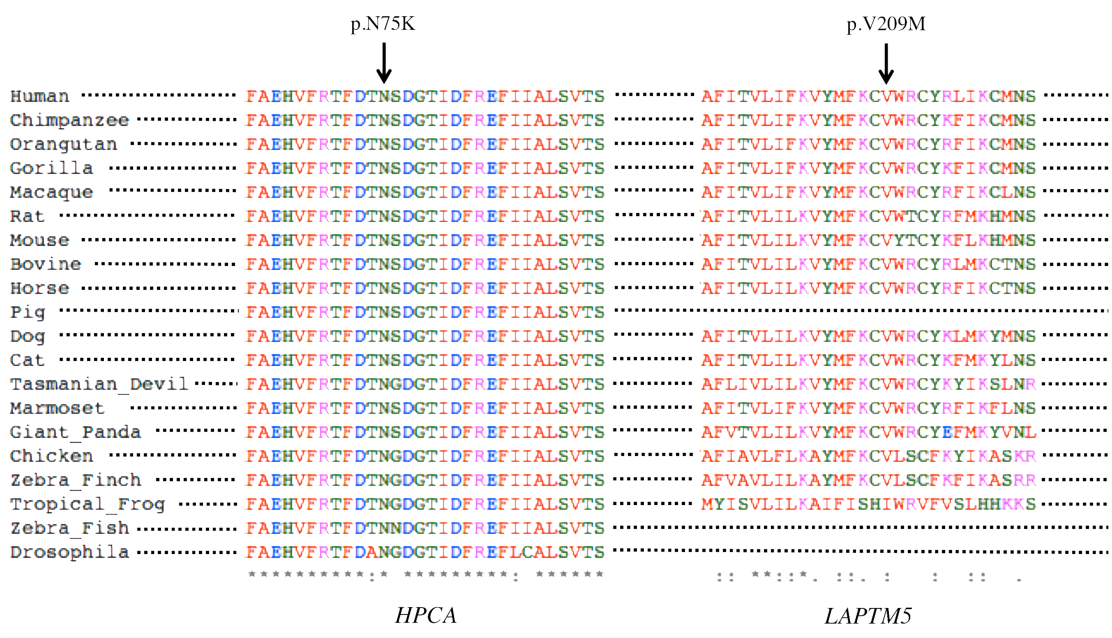

Figure 22 - Orthologous protein sequences for the relevant variant-containing regions *HPCA* and *LAPTM5*, obtained via Uniprot and Flybase for all species available and aligned using ClustalOmega. No orthologous sequence was available for worm (*C. Elegans*) in either case. The affected amino acid in the p.N75K mutation in *HPCA* (left) shows absolute interspecies conservation. Note also the high level of conservation in this region as a whole. The amino acid affected by the p.V209M mutation in *LAPTM5* (right) is not fully conserved, with the tropical frog possessing an isoleucine at this position, and no orthologue exists for pig, zebra fish or drosophila. Symbols under each column indicate the degree of conservation (an asterisk = a single identical amino acid; a colon = strongly similar properties; a period = weakly similar properties; a blank = no conservation]. Colours indicate physiochemical properties of amino acids (red = small/hydrophobic; blue = acidic; magenta = basic; green = hydroxyl/sulfhydrl/amine/glycine).

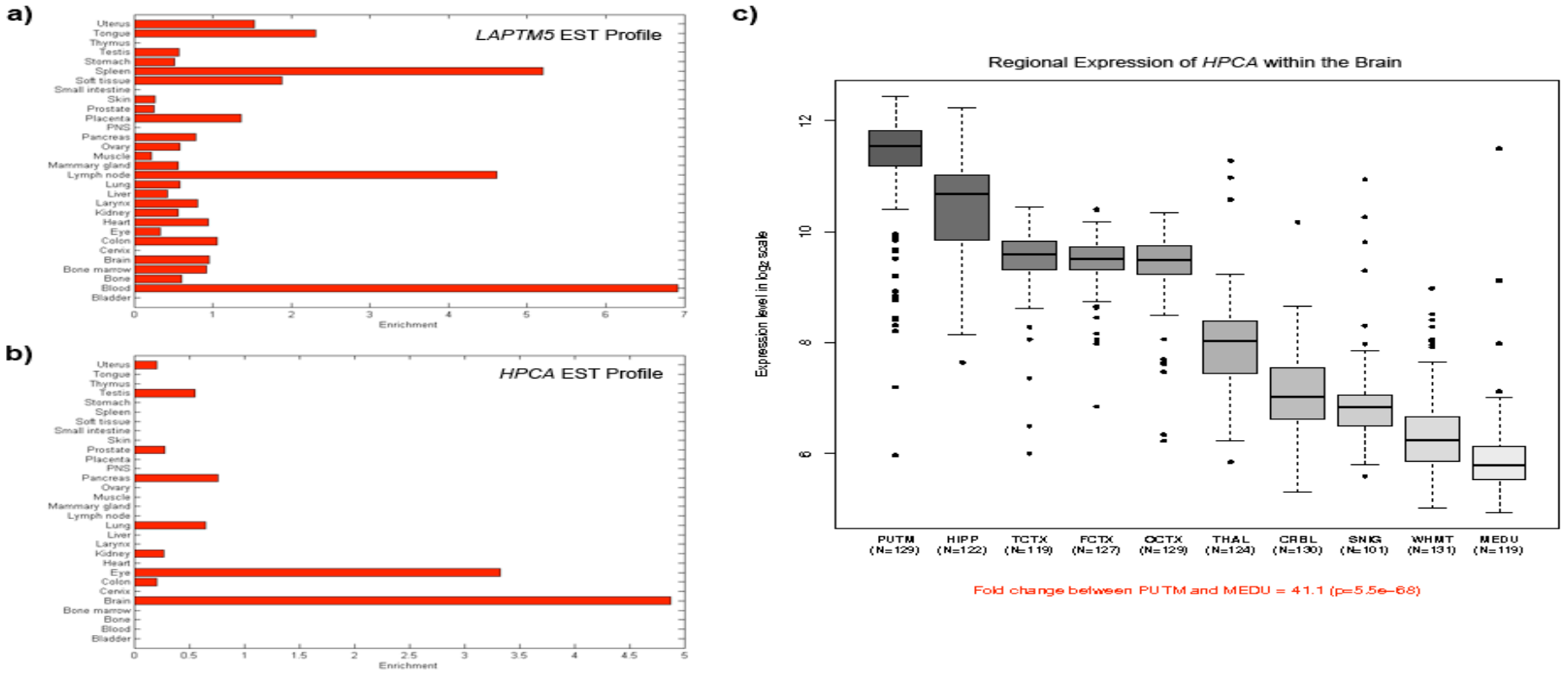

Figure 23 - Publically-available expressed sequence tag data for a) *LAPTM5* and b) *HPCA*, demonstrating that both genes show relatively tissue-specific expression patterns. *LAPTM5* is predominantly expressed in hemopoietic tissues, whilst *HPCA* is almost exclusively expressed in the brain. c) Box plot of mRNA expression levels for *HPCA* in 12 CNS regions, based on exon array experiments and plotted on a log2 scale (y axis). This plot shows significant variation in *HPCA* transcript expression across the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). *HPCA* mRNA expression is highest in the putamen, followed closely by the hippocampus. Expression is also high in the cortex. Whiskers extend from the box to 1.5 times the inter-quartile range.

Furthermore, our own in-house expression datasets based on exon array profiling of samples from multiple areas of ~130 control brains not only confirm that *HPCA* is well expressed in the brain as a whole, but also demonstrates significant regional differences in expression, with the highest levels of expression being detected in the striatum (represented by the putamen in our study), an area intimately connected with the pathophysiology of many movement disorders, including dystonia (Fig 23C). Indeed, expression of *HPCA* was detected at levels 41 fold higher in the striatum than in the medulla ($p=5.5x10^{-68}$), the region with the lowest levels of expression relatively-speaking, even though absolute levels of *HPCA* expression in the medulla are still quite high. Expression of *HPCA* is also high in the hippocampus and cortex.

7.3.4 Mutational Screening of Both Candidate Variants in an Independent Cohort of Young Onset Dystonia

In an attempt to find a second novel mutation in either gene, we next sequenced the relevant exons of both genes in 150 cases of young onset dystonia, as well as a DNA sample from an affected member of the DYT17 kindred (see section 7.2.2 for full case selection criteria). We did not detect any further novel variants in exon 7 of *LAPTM5*. In exon 2 of *HPCA*, however, we detected a second, heterozygous, novel, missense variant (c.212C>A) resulting in an amino acid substitution (p.T71N) at a position in the protein, just 4 amino acids before the location of original homozygous mutation found in the index family. The affected nucleotide shows extremely high conservation scores (PhyloP=5.76 [max=6]; PhastCons=1 [max=1]). Although the amino acid at position 71 is not itself recognized as an obligatory Ca^{2+}-coordinator, it is still within the second EF-hand domain (amino acids 60 to 95, according to Uniprot) and does show the same absolute interspecies conservation as the amino acid affected by the N75K mutation. In addition, it is predicted to be damaging by all 4 *in silico* prediction programs. In view of this finding, we went on to sequence the remainder of the coding exons of *HPCA* in the same 151 samples. We found only one additional variant of any kind and this was in the same sample that harboured the p.T71N mutation, meaning that individual was compound heterozygous for these two changes. The additional variant was a novel missense mutation (c.568G>C; p.A190T) located towards the end of exon 4 of *HPCA*, which encodes the C-terminal of the protein. The nucleotide

involved is conserved (PhyloP=2.015; PhastCons = 1), whilst the affected amino acid is conserved in most species, with the exception the fly (*D. melanogaster*) and the Tasmanian devil (*S. harrisii*). However, only MutationTaster predicts it to be disease-causing; SIFT, Provean and PolyPhen2 predict that the substitution will be tolerated.

Having detected two potentially pathogenic mutations in *HPCA* in the same DNA sample of an individual with dystonia, we established contact with the patient, a 64 year-old woman of Sri Lankan origin, to verify the medical and family history and perform a full neurological examination. She reported the onset of dystonia in her early twenties, possibly even her late teens, which initially manifested with abnormal involuntary movements of her fingers that were most noticeable while trying to type. Over time, her dystonia has very gradually worsened, with the emergence of a tremulous component. Nonetheless, it remains segmental, affecting only the hands, arms and muscles of the neck, and would be classified clinically as mild. Despite being a member of sibship of seven, no other family member, including her parents, siblings, and her siblings' children, reported any symptoms consistent with dystonia (figure 24). As both parents and two of her siblings were now deceased, we were only able to obtain DNA samples from the 4 surviving siblings, which we used to sequence the relevant exons of *HPCA* for segregation analysis. The results of the segregation analysis in this family were consistent with the hypothesis that the compound heterozygous mutations we observed in *HPCA* in this individual are the cause of her dystonia, in that 1) one sibling was heterozygous for the p.A190T mutation alone, thus confirming that the two mutations had been inherited separately from each parent; and 2) the only affected individual was also the only individual in which we detected both the p.T71N and p.A190T mutations in the compound heterozygous state; and 3) the remaining unaffected siblings all possessed either one or both wild type alleles, incompatible with development of a recessive disease (figure 24).

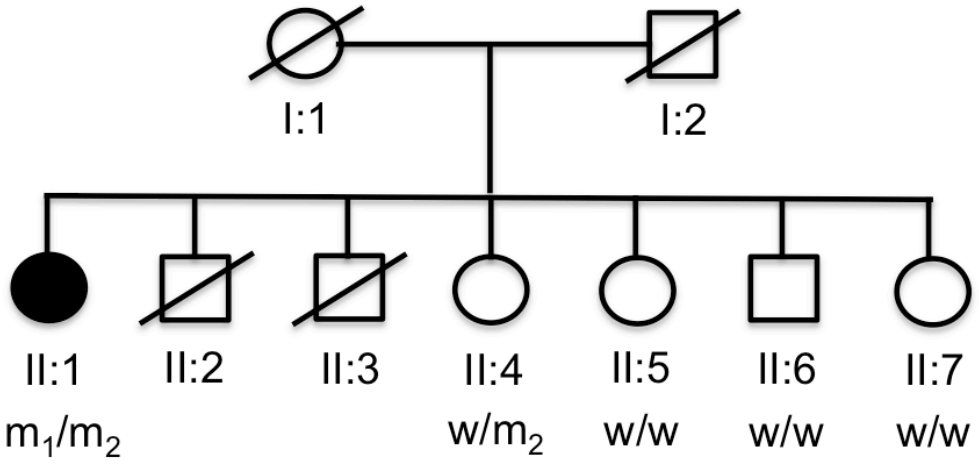

Figure 24 – Genetic pedigree of the second family with compound heterozygous mutations in *HPCA*. Individual II:1 developed young onset dystonia affecting the hands in her late teens or early twenties. It has gradually extended to become segmental, involving the hands, arms and neck. No other family member was affected. Mutational status is shown under each individual: w = wildtype allele; m1 = p.T71N mutation; m2 = p.A190T mutation.

7.3.5 Further Mutational Screening in an Additional 288 Cervical/Upper Limb Onset Predominant Dystonia Cases

We next attempted to find a third case by screening all 3 coding exons of *HPCA* in a second cohort of 288, non-autosomal-dominant cases of young-onset (<40 years of age), isolated dystonia in which the dystonia was either generalized or most prominent in the upper limbs, cervical or cranial region. Unfortunately, no further variants of any kind – either novel or previously annotated – were detected in any sample.

7.3.6 Low Level of Natural Human Sequence Variation in HPCA

Given the failure to find even a single sequence variant of any sort in 438 individuals, we decided to examine the burden of natural human genetic variation in *HPCA*. Using the pooled next-generation sequencing results of the NHLBI Exome Sequencing, ClinSeq, 1000 Genomes, and HapMap projects, which, contain collectively the exome sequencing data from 8,451 individuals of various ethnicities, we identified only 8 separate missense variants (none of them homozygous) in this gene (Table 22). By way

of comparison, based on the same datasets, 31 missense (11 of which are predicted to be damaging by both SIFT and PolyPhen), 2 splice site and 1 frameshift variant have been detected in *LAPTM5*. Although, due to a lack of data, we cannot be certain this low level of variation in *HPCA* extends to all populations, this observation does at least further increase the likelihood that the identification of compound heterozygous mutations in *HPCA* in a second dystonia kindred, segregating perfectly with disease, was very unlikely to have occurred by chance.

7.3.7 Neuropsychological Testing in an Affected Member of the Index Family

Given that the protein product of *HPCA* has been implicated in synaptic plasticity and that knock-out mice for this gene show defects in spatial and associative learning, we arranged for detailed neuropsychological testing of an affected member of the index family by a trained clinical psychologist who was blinded to the purpose of the assessment. The results revealed a degree of cognitive under-functioning with respect to optimal premorbid estimates. In particular, there was difficulty encoding visual and verbal information, difficulty with attention and processing speed, and mild problems with executive dysfunction. Delayed visual and verbal recall memory, naming, visual constructional abilities and phonemic and category fluency were intact. Although considerable anxiety was noted during testing, the clinical psychologist was of the opinion that this may have exacerbated but did not fully explain, the cognitive deficits evident on testing.

Table 22 – Publically annotated variants with protein-coding consequence in human *HPCA*. In total, this represented a pool of 8,451 individuals (16,902 chromosomes). Only 8 separate missense variants were detected, none of which were common and none of which were homozygous. Only two were predicted to be damaging by both SIFT and PolyPhen.

SNP Identifier	c.DNA change (NM_002143.2)	Protein Change (NP_002134.2)	Database of Origin	Genotype Count (/8,451)	SIFT Prediction	PolyPhen Prediction
rs11554958	c.63G>T	p.Glu21Asp	HapMap	G/G = 8,450 G/T = 1 T/T = 0	Benign	Benign
rs147332564	c.178G>C	p.Asp60His	NHLBI ESP	G/G = 8,450 G/C = 1 C/C = 0	Damaging	Damaging
rs182483890	c.196G>A	p.Glu66Lys	1000 Genomes	G/G = 8,450 G/A = 1 A/A = 0	Benign	Damaging
rs201850746	c.286C>T	p.Arg96Cys	ClinSeq	C/C = 8,450 C/T = 1 T/T = 0	Damaging	Benign
rs138767632	c.373G>A	p.Val125Met	NHLBI ESP	G/G = 8,449 G/A = 2 A/A = 0	Damaging	Damaging
rs376349097	c.403G>T	p.Val135Leu	NHLBI ESP	G/G = 8,450 G/T = 1 T/T = 0	Benign	Benign
rs140440243	c.412A>C	p.Met138Leu	NHLBI ESP	A/A = 8,450 A/C = 1 C/C = 0	Benign	Benign
rs371851892	c.484G>A	p.Gly162Ser	NHLBI ESP	G/G = 8,450 G/A = 1 A/A = 0	Benign	Benign

7.3.8 Functional Studies in Rat Neuron Primary Culture

Hippocalcin, the protein product of *HPCA*, is a neuronal calcium sensor protein, believed to play an important role in intracellular calcium-dependent signalling. The N75K mutation that we detected in the index family cause the substitution of a key, calcium-coordinating amino acid that forms part of the almost invariable, universal sequence motif that defines all functioning EF-hand domains (only asparagine or aspartic acid are found at this position; see also discussion below). The loss of this motif and, in particular, the substitution of a positively charged lysine at an amino acid position directly involved in the binding of a positively charged Ca^{2+}, would be expected to reduce the domain's affinity for the ion or even block its binding completely. Given that calcium-binding is known to be a prerequisite for the conformational change essential to the protein's function, we hypothesized that the N75K mutation at least would be expected to result in significantly reduced or even complete loss of function of hippocalcin within the cell. Therefore, we performed shRNA knockdown of *hpca* on rat primary neuronal and astrocytes and used the resultant cells to measure, by means of fura-2 florescence microscopy, the effect of hippocalcin deficiency on cellular calcium homeostasis after exposure to different neuropharmacological agents. Specifically, cells were stimulated with: glutamate (10 μM) to simulate a physiological calcium signal in neurons via activation of glutamate receptors; ATP (100μM) to stimulation of P2Y receptors in astrocytes; and potassium chloride (KCl, 50mM) to depolarize neuronal membrane and induce opening of voltage-gated calcium channels.

Unfortunately, it is necessary to state at this point that the following results need to be interpreted with caution. As is normal practice, *hpca* knock-down is transfected rodent neurons was confirmed by Western blot. Regrettably, after publication of the data, the blots themselves have been misplaced and cannot therefore be included in this thesis as proof of successful knock down. I have decided nonetheless to include the data as it is interesting, but thankfully is not required as part of the genetic proof for the causality of this gene.

We found that application of glutamate to *hpca*-shRNA transfected neurons resulted in smaller, but non-significant neuronal response to glutamate (0.95±0.4, n=35; see figure

25A, lower panel), compared to the scrambled and empty vector transfected controls (1.35±0.4, n=35, p=0.48 and 1.45±0.3, n=21, p=0.38, respectively; see figure 25B, upper panel).

Astrocytic signal to stimulation with ATP (100 μM) in *HPCA* knock-down cells was also smaller (figure 25C, D; 0.6±0.2, n=55) than those in scrambled and empty vector transfected control cells (1.4±0.35, n=45 (scr) and 0.9±0.5, n=28, respectively), but the reduction in signal only reached significance when compared to scrambled controls (p=0.04).

Somewhat unexpectedly, the most significant difference between *hpca*-silenced and control neurons was observed after depolarization of the plasma membrane by application of 50 mM KCL. As anticipated, in both scrambled and empty vector transfected control neurons, this resulted in a strong Ca^{2+}-signal (1.65±0.45, n=35 and 1.75±0.34, n=21, respectively). In *hpca*-shRNA transfected neurons, however, application of KCL produced almost no observable calcium signal (0.1±0.02, n=35; p=0.001 vs scrambled and p<0.0001 vs empty vector controls; Figure 25A and B). Importantly, all of these cells demonstrated a clear response to glutamate confirming their neuronal origin (figure 25B).

Together, this pattern of severely altered neuronal responses to physiological stimuli suggests that *HPCA* deficiency may result in inhibition of potential-sensitive Ca^{2+} channels or, alternatively, modify the mechanism of maintaining the membrane potential, affecting cellular response to membrane depolarization. As stated above, however, these results are included for interest's sake only and, in the absence of the Western blot confirming successful knock-down, cannot form part of the argument for the causality of this gene in this kindred.

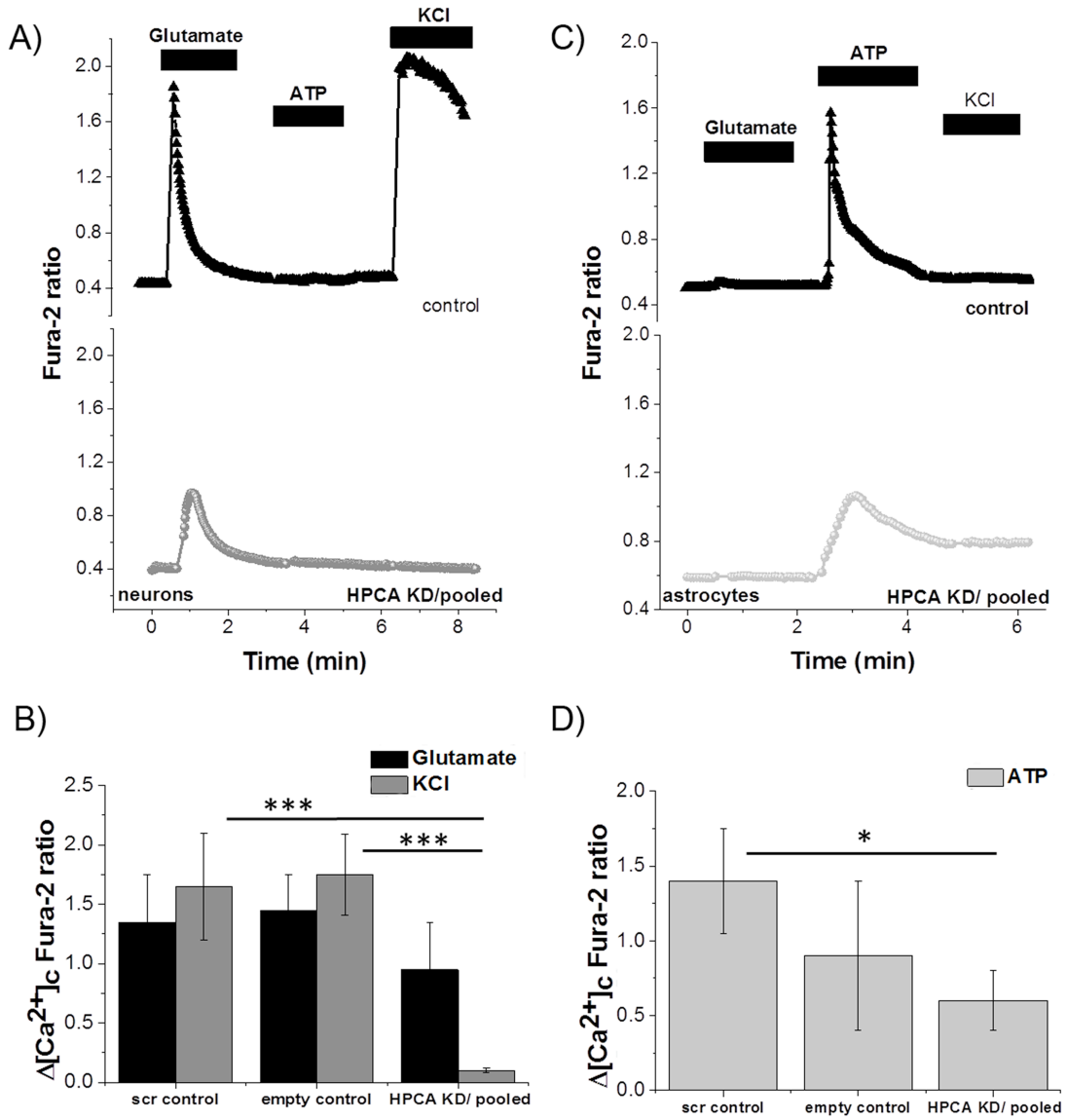

Figure 25 – A and C) Representative traces of response of $[Ca^{2+}]_c$ to physiological stimuli as measured by changes in Fura-2 fluorescence intensity. *hpca* knockdown neurons show no rise in $[Ca^{2+}]_c$ in response to depolarisation of the plasma membrane with 50 mM KCl (A, dark grey trace; B, dark grey bars) in comparison to neurons from the scrambled or the empty control (A, black triangle trace). In addition, the amplitude of the response to physiological concentration of Glutamate (5μM) in the *hpca* knockdown neurons was decreased compared to the control (b, black bars), although not statistically significant. *hpca* knockdown also diminished the amplitude of the $[Ca^{2+}]_c$ response of astrocytes to an ATP stimulus (100μM) (C, D - light grey trace and bars). Error bars represent SEM; asterisks represents statistical significance (*** = $p<0.0001$, * = $p<0.05$).

7.4 Discussion

We set out to use homozygosity mapping and whole exome sequencing to help elucidate the genetic cause of the young-onset, isolated dystonia in a consanguineous kindred, suggesting autosomal recessive inheritance. By appropriate filtration of the variants, we were able to narrow our list of candidates down to just two potentially causal mutations in the genes *LAPTM5* and *HPCA*, respectively. Given the neurological nature of the disease under consideration, expression data suggested *HPCA* to be the more probable candidate as it is brain-specific and is detected at the highest levels in the striatum, an area of the brain often implicated in movement disorders. After sequencing the exon harboring the candidate variants for both genes in an independent cohort of young-onset dystonia, we identified a second kindred with isolated dystonia that segregated with novel, compound heterozygous mutations in *HPCA*, suggesting recessive mutations in this gene to be the cause of the isolated dystonia in both this and the original index family. Despite further sequencing of the whole gene in both this cohort plus an additional case, we did not identify any cases with recessive mutations in *HPCA*. In fact, we did not find any other coding or splice site variant in *HPCA* – either novel or previously annotated – in any of the other cases screened. This low level of variation is confirmed by the results of the NHLBI exome sequencing project, which has detected only 6 missense variants (none of them homozygous) and 12 synonymous coding variants in *HPCA* in the 13,000 chromosomes for which data has so far been made public (European and African American populations combined). Although we cannot be certain this low level of variation extends to all populations because of a lack of data, this observation does at least further increase the likelihood that the identification of compound heterozygous mutations in *HPCA* in a second dystonia kindred, segregating perfectly with disease, was very unlikely to have occurred by chance.

Notwithstanding the loss of the Western blot, functional studies in rodent neurons and astrocytes, knocked-down for *HPCA* gene, suggest that *HPCA* deficiency may modify the physiological calcium signal in both neurons and astrocytes. Most importantly, we found that hippocalcin regulates the two most common pathways for calcium entry – glutamatergic receptors and potential-sensitive calcium channels. This might suggest a

role of hippocalcin in maintenance of the plasmalemmal membrane potential either by affecting it indirectly or by direct influence on the voltage dependent calcium channels. Further experiemental work is required to confirm this, however.

Hippocalcin is part of a subfamily of neuronal calcium sensor (NCS) proteins with high sequence homology, which includes vasinin-like proteins 1 and 2, hippocalcin-like protein 1 and neurocalcin δ (figure 26) [342]. All NCS proteins are characterized by four EF-hand domains that act as potential Ca^{2+} binding sites. EF-hand domain 1, however, is invariably inactive, whilst EF-hand domain 4 is active in some, but not all NCS proteins. Canonical calcium-binding EF-hand domains are characterized by the semi-conserved sequence motif **D**-X-**D**/**N**-X-**D**/**N**-G-$(X)_5$-**E**, where the obligate amino acids in bold are involved in the coordinative binding of Ca^{2+} [342, 343].

```
Hippocalcin                 MGKQNSKLRPEMLQDLRENTEFSELELQEWYKGFLKDCPTGILNVDEFKKIYANFFPYGD   60
Neurocalcin_Delta           MGKQNSKLRPEVMQDLLESTDFTEHEIQEWYKGFLRDCPSGHLSMEEFKKIYGNFFPYGD   60
Hippocalcin-Like_Protein_1  MGKQNSKLRPEVLQDLRENTEFTDHELQEWYKGFLKDCPTGHLTVDEFKKIYANFFPYGD   60
Visinin-Like_Protein_1      MGKQNSKLAPEVMEDLVKSTEFNEHELKQWYKGFLKDCPSGRLNLEEFQQLYVKFFPYGD   60
Hippocalcin-Like_Protein_4  MGKTNSKLAPEVLEDLVQNTEFSEQELKQWYKGFLKDCPSGILNLEEFQQLYIKFFPYGD   60
                            *** **** **:::** :.*:*.: *:::*******:***:* *.::**:::* :******

Hippocalcin                 ASKFAEHVFRTFDTNSDGTIDFREFIIALSVTSRGRLEQKLMWAFSMYDLDGNGYISREE  120
Neurocalcin_Delta           ASKFAEHVFRTFDANGDGTIDFREFIIALSVTSRGKLEQKLKWAFSMYDLDGNGYISKAE  120
Hippocalcin-Like_Protein_1  ASKFAEHVFRTFDTNGDGTIDFREFIIALSVTSRGKLEQKLKWAFSMYDLDGNGYISRSE  120
Visinin-Like_Protein_1      ASKFAQHAFRTFDKNGDGTIDFREFICALSITSRGSFEQKLNWAFNMYDLDGDGKITRVE  120
Hippocalcin-Like_Protein_4  ASKFAQHAFRTFDKNGDGTIDFREFICALSVTSRGSFEQKLNWAFEMYDLDGDGRITRLE  120
                            *****:*.***** *.********** ***:**** :**** ***.******:* *:: *

Hippocalcin                 MLEIVQAIYKMVSSV--MKMPEDESTPEKRTEKIFRQMDTNNDGKLSLEEFIRGAKSDPS  178
Neurocalcin_Delta           MLEIVQAIYKMVSSV--MKMPEDESTPEKRTEKIFRQMDTNRDGKLSLEEFIRGAKSDPS  178
Hippocalcin-Like_Protein_1  MLEIVQAIYKMVSSV--MKMPEDESTPEKRTDKIFRQMDTNNDGKLSLEEFIRGAKSDPS  178
Visinin-Like_Protein_1      MLEIIEAIYKMVGTVIMMKMNEDGLTPEQRVDKIFSKMDKNKDDQITLDEFKEAAKSDPS  180
Hippocalcin-Like_Protein_4  MLEIIEAIYKMVGTVIMMRMNQDGLTPQQRVDKIFKKMDQDKDDQITLEEFKEAAKSDPS  180
                            ****::******.:*  *:* :*  **::*.:*** :** :.* :::*:** ..******

Hippocalcin                 IVRLLQCDPSSASQF 193
Neurocalcin_Delta           IVRLLQCDPSSAGQF 193
Hippocalcin-Like_Protein_1  IVRLLQCDPSSASQF 193
Visinin-Like_Protein_1      IVLLLQCDIQK---- 191
Hippocalcin-Like_Protein_4  IVLLLQCDMQK---- 191
                            ** ***** ..
```

Figure 26 - Sequence alignments for the subfamily of human NCS proteins of which hippocalcin is a member. Protein sequences demonstrate a high degree of sequence homology, particularly between hippocalcin, neurocalcin delta and hippocalcin-like protein 1. The area of shading at the N-terminal region of the protein indicates the myristoyl moiety, whilst the following three areas of shading indicate the potentially functional EF-hand domains 2, 3 and 4. EF hand domain 1 is non-functional in all members of the family and is not shaded in this figure. Colours indicate physiochemical properties of amino acids (red = small/hydrophobic; blue = acidic; magenta = basic; green = hydroxyl/sulfhydrl/amine/glycine).

The binding of calcium to the EF hand domains of NCS proteins operates a myristoyl-switch mechanism that controls the protein's ability to translocate to target membranes and/or interact with downstream effectors (figure 27) [344-346]. This mechanism has been extensively studied in the related neuronal protein, recoverin [346]. In the native, Ca^{2+}-free state, recoverin and other such myristoyl-switch proteins assume a physical configuration that sequesters the myristoyl-side chain in an internal hydrophobic pocket. After a rise in local Ca^{2+} concentrations, the functional EF-hand domains are believed to bind Ca^{2+} in sequential, semi-cooperative manner [347]. With the binding of the final Ca^{2+} ion, there is a conformational change that exposes the hydrophobic parts of the protein and causes the extrusion the N-terminal myristoyl side chain, freeing it for interaction with cellular membranes and/or target proteins. In this way, Ca^{2+}-myristoyl switch proteins are able to act as reversible transducers of cellular Ca^{2+} signals, capable of integrating both temporal and spatial aspects of signaling over a tight dynamic range [344].

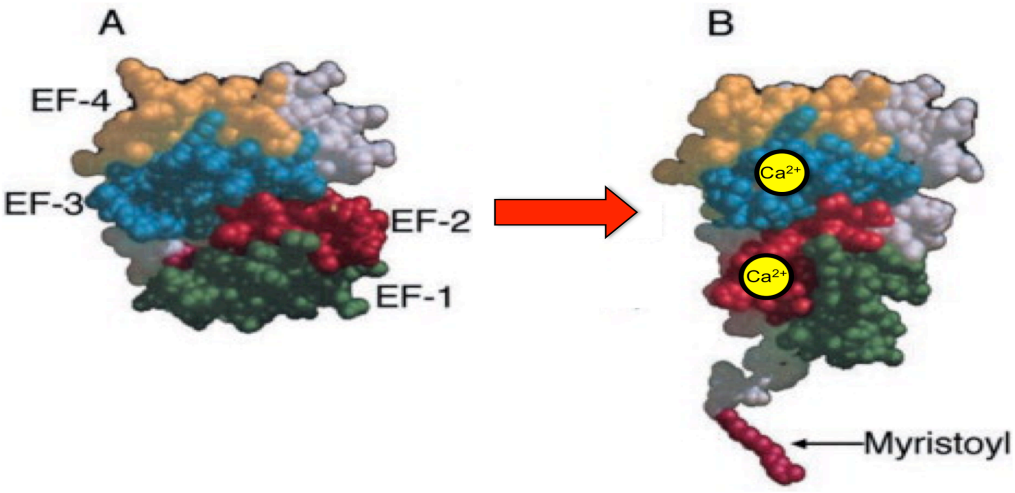

Figure 27 - A semi schematic diagram of the calcium-myristoyl switch mechanism using recoverin, a related neuronal calcium sensor protein as an example. A) In the native, non-Ca^{2}+ bound state, recoverin assumes a configuration in which the myristoyl moiety (just visible in magenta) is buried deep within a hydrophobic pocket within the protein. B) Upon binding of Ca^{2+} first to EF-hand domain 3 and then to EF-hand domain 2, the protein undergoes a conformational change that causes the myristoyl moiety to be extruded, freeing it for interaction with target membranes and proteins. (This figure has been modified from Figure 1 as published by Ames et al., 1997)

From the point of view of the molecular mechanism mediating the pathogenicity of the mutations we detected, the p.N75K mutation in *HPCA* results in a substitution of the second Ca $^{2+}$-coordinating residue of the binding sequence within EF-hand domain 2. As detailed above, in functional EF-hand domains, usually only one of two amino acids will be found at this position: a negatively charged aspartic acid or a neutral asparagine (as in hippocalcin). The p.N75K mutation thus not only results in the loss of a canonical calcium-coordinating amino acid, but also results in the incorporation of a positively-charged lysine, which might reasonably be expected to present a particular impediment to the binding of the similarly-charged Ca^{2+} ion. Thus, it is plausible to hypothesize that the homozygous p.N75K mutation, by impairing or even preventing calcium binding to EF-hand domain 2, might be expected to reduce the likelihood of the any conformational change of hippocalcin in response to Ca^{2+} signals. Since conformation change is a prerequisite for its interaction with downstream effectors, the ultimate outcome would be a defect in cellular Ca^{2+} signal transduction. In the case of the second family we identified, the p.T71N mutation is also located in EF-hand domain 2 and, despite not affecting an amino acid directly involved in coordinating Ca^{2+}, could potentially adversely affect binding kinetics via alterations of local secondary or tertiary structure. The p.A190T mutation, on the other hand, is not located in any EF-hand domain but, instead, in the C-terminal region of the protein. The mechanism by which this mutation might impair protein function is less obvious. Nonetheless, it has been suggested that the C-terminal regions of other NCS proteins may be involved in fine-tuning of their response or the determination of target specificity [348, 349]. Follow-up studies involving direct mutagenesis will help definitively answer these mechanistic questions.

Neuronal calcium sensor proteins, such as hippocalcin, possess no inherent enzymatic properties, but exert their calcium-dependent functions through interactions with other proteins. Thus, the identification of downstream interactors is a necessary first step in understanding their specific activities. In the case of hippocalcin, the identity of these interactors remains incomplete. There is, nonetheless, evidence to suggest that it may have a role in 1) the modulation of cyclic nucleotide signaling in the olfactory epithelium [350]; 2) long-term depression in the hippocampus [351-353]; 3) generation of the

slow afterhyperpolarization current, important is controlling neuronal excitability [354-356]; 4) regulation of gene transcription [357, 358]; and 5) neurite outgrowth [359]. Whilst it is entirely possible that some other, currently unknown function of hippocalcin may underlie its involvement in dystonia, it notable that, in relation to the processes mentioned above, both the aberrant excitability of striatal neurons and altered synaptic plasticity due in part to decreased long-term depression are two mechanisms believed to be important in at least some forms of dystonia [360-363]

Hippocalcin has been most intensely studied in relation to its role in synaptic plasticity within the hippocampus where it has been hypothesized to play a role in memory formation. Indeed, hippocalcin knock-out mice showed deficits in tests of spatial and associative memory, in the absence of any obvious structural abnormalities within the brain [357]. Difficulties with memory were not, however, reported by any of the individuals with hippocalcin mutations in our study. Although the aim of this study was to identify the genetic cause of dystonia in the index family and not to clinically re-phenotype them, we were able to arrange neuropsychological testing for one of the affected members of the index family who travels to our center frequently. This did show evidence of cognitive underfunctioning, with particular difficulties in encoding verbal and visual information. The affected individual in the second family – who also had a much milder dystonic phenotype – had only just recently retired as a teacher and had an MMSE of 30/30, arguing against any significant difficulties with memory or learning and limitations in the studies ethical approval precluded further testing. Whether memory deficits represent subtle and variable associated phenotype in humans with *HPCA* mutations (like hyposmia in *GNAL* mutations [364]) or whether the neuropsychological deficits observed in the individual in this study were an incidental finding remains an open question that only the identification of further cases and a study dedicated to their neuropsychological profiling will answer.

In summary, in this chapter I have presented evidence to support biallelic mutations in *HPCA* as the first identified cause of autosomal recessive, isolated dystonia. This discovery permits the assignation of an actual causal gene to the anomalous DTY2 locus and demonstrates the existence of a recessive form of isolated dystonia. Although we

were only able to identify one additional family with dystonia presumed secondary to mutations in *HPCA*, this is not inconsistent with the relative rarity of the disorder and it is unlikely that any European or American institution would have possesed a significantly higher number of cases. However, screening of further suitable cases, particularly in geographical regions with higher rates of consanguinity, will be required to give a better idea of its actual prevalence.

CHAPTER 8:

Exome Sequencing in Autosomal Recessive Cervical-Onset, Dopa-Responsive Dystonia

8. Exome Sequencing in Autosomal Recessive, Cervical-Onset Dopa-Responisve Dystonia

8.1 Introduction

Dopa-responsive dystonia is an uncommon condition, with a prevalence of 0.5 – 1 individual per million. Nonetheless, it remains an important condition to recognise as it is, by definition, readily treatable by the oral administration of L-dopa. Response to this treatment is often dramatic, sustained and is not usually associated with the development of motor fluctuations and dyskinesias, which tend to develop with prolonged treatment in other dopa-responsive disorders, most notably Parkinson's disease. To date, 3 genes have been convincingly shown to cause dopa-responsive dystonia: *GCH1* (GTP Cyclohydrolase 1), *TH* (Tyrosine Hydroxylase) and *SPR* (Sepiapterin Reductase) [234-236]. All three genes encode enzymes that are involved in the endogenous biosynthesis of dopamine.

We identified a Muslim Indian kindred in which three out of five siblings presented with dopa-responsive, predominantly cervical dystonia. Previous mutational screening had ruled out *GCH1*, *TH* or *SPR* as the cause of the disorder [365]. Using a combination of linkage analysis and exome sequencing, we sought to identify the cause of the dopa-responsive dystonia in this family.

8.2 Subjects, Materials and Methods

8.2.1 Clinical Details of the Index Family

Participants were drawn from a multi-generational Muslim Indian family. The core pedigree of the family as initially provided to us is shown in figure 28. No other member of the extended family was said to have suffered from any symptoms that might be compatible with dystonia. DNA was initially available from three affected individuals as well as both parents. The pattern of inheritance was consistent with an autosomal recessive condition. The parents reported no consanguinity.

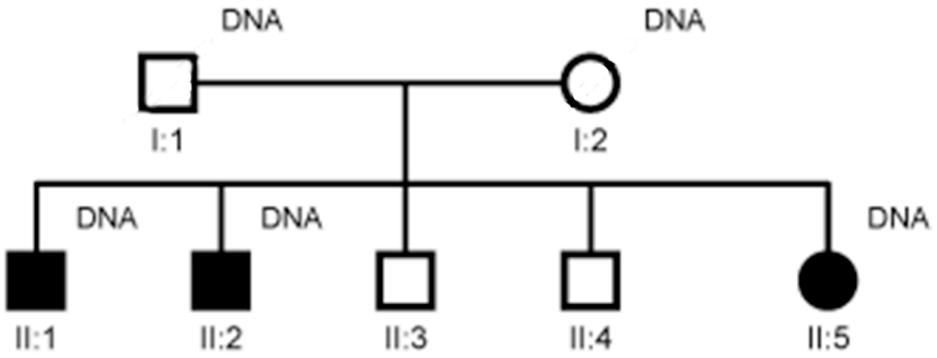

Figure 28 – Pedigree as initially provided for the core index family. DNA is marked as was available. All affected siblings demonstrated dopa-responsive cervical dystonia. No member of the extended pedigree reported or were reported to have symptoms consistent with dystonia.

Individual II-1 developed jerky right-sided torticollis at the age of 15, which progressively worsened over the period of a year. Subsequently, laryngeal dystonia, clawing of his fingers, hand tremor, and difficulty writing also developed. On examination at 18 years of age, he had right torticollis, left sternocleidomastoid hypertrophy, head and hand tremor, and tremulous writing. Distal finger contractures, mild right scoliosis, and spasmodic dysphonia were noted. Treatment with trihexyphenidyl and baclofen did not produce any benefit. However, on initiation of treatment with levodopa-carbidopa (100:25mg three times daily), there was dramatic improvement in his cervical dystonia and he was able to resume work. The postural hand tremor, however, remained unchanged.

Individual II-2 and II-3 were similarly affected, developing cervical dystonia with associated side-to-side head tremor at ages 13 and 11 years respectively. In addition, there was postural hand tremor that interfered with writing. In the case of both additional siblings, treatment with levodopa-carbidopa at low doses produced significant improvement in the degree of cervical dystonia.

No progression of dystonia occurred on follow-up for 12, 6, and 5 years in the three siblings. The excellent response of the dystonia to low doses of levodopa was sustained in all three cases. Withdrawal of levodopa resulted in the return of symptoms within 4 days and these resolved again after levodopa was readministered.

8.2.2 Whole Exome Sequencing

DNA was extracted from whole blood samples obtained from all five siblings and both parents. Whole exome sequencing was performed as per section 4.15. Reads were subsequently aligned and annotated using the standard approaches set out in section 4.16.

Initially DNA from individual II-1 only was used for exome sequencing. However, this left a too large number of potential candidate variants for consideration aand so, subsequently, DNA from both other affected individuals were also exome sequenced. Variant filtration is described in the results section 8.3.4.

8.2.3 Genotyping and Autozygosity Mapping

The family did not report any consanguinity and we hypothesised that disease in this family was most likely secondary to compound heterozygous mutations. However, we undertook autozygosity mapping, as per section 4.13, in any case to quantify the extent of any tracts of homozygosity and thus provide a estimation of the likelihood of unreported consanguinity. Genome-wide SNP data was obtained for all three affected individuals using the OmniExpress, which analyses approximately 500,000 markers, as per section 4.12.

8.2.4 CNV Analysis

Genome-wide SNP data was also used to search for shared deletions and duplications that might be the cause of the dopa-responsive dystonia in this family. Potential CNVs were detected using PennCNV for each sample individually. Subsequently, CNVs that were common to all three affected individuals were identified.

8.2.5 Linkage Analysis

Linkage analysis was performed using MERLIN as per section 4.14. A parametric analysis under an autosomal recessive model of disease inheritance with penetrance of 100% and a pathological allele frequency of 0.0001 produced the best scores.

8.3 Results

8.3.1 Autozygosity Mapping

Autozygosity mapping demonstrated that shared areas of homozygosity in this family were both rare and small (see table 23 for a comparison of the size and distribution of homozygosity in a typical consanguineous kindred). In assition, no potentially pathogenic, homozygous mutations were found in these regions. Based on this information, we concluded that the report of no consanguinity in the core kindred was correct and that the dystonia was most likely the result of compound heterozygous mutations.

Table 23 - Areas of extended (>1 megabase) homozygosity shared by all the affected siblings with genomic coordinates, length (in kilobases) and number of genes covered.

Chr	Start Position	Stop Position	Length (Kb)	Genes	Potential Pathogenic
3	48636099	49860854	1224.76	49	0
3	50368806	51887557	1518.75	27	0
7	118356108	120086499	1730.39	3	0
8	99413330	100993147	1579.82	6	0
12	60601848	61732540	1130.69	3	0

8.3.2 Linkage Analysis

Linkage analysis in a kindred of this size was not expected to produce a truly significant LOD score, but was, instead, used to highlight areas with a greater likelihood of harboring a causal variant and identify others where potentially causal variants could be

excluded (i.e. areas with LOD score of less than -2). The summary of the genome wide linkage data is shown in figure 29 and table 24. The highest LOD score achieved was 1.204 on chromosome 6, 11 and 12, which defined the primary areas of interest. These peaks were also the widest of the linkage peaks, suggesting a large common haplotype. There were a further 5 peaks with a maximal LOD score of between 1 and 1.203, which defined the secondary areas of interest.

Table 24 - Regions of linkage with a LOD score of greater than 1. Three peaks had an identical LOD score of 1.204 and are identified as the primary areas of interest. For each region, the genomic coordinates, size, LOD score and number of known genes within the region are given (hg19 via Ensembl),

Chr	Start Position	Stop Position	Size (Mb)	Max LOD	Known Genes
PRIMARY AREAS OF INTEREST					
6	506231	42044945	41.5	1.204	364
11	76996593	114753038	37.8	1.204	407
12	2456115	48272895	45.8	1.204	548
SECONDARY AREAS OF INTEREST					
3	43797458	65180001	21.4	1.203	318
3	172843082	187269748	14.4	1.181	149
6	100785588	127726196	26.9	1.202	208
7	49733584	77977855	28.2	1.116	423
8	31159012	60237300	29.1	1.180	240

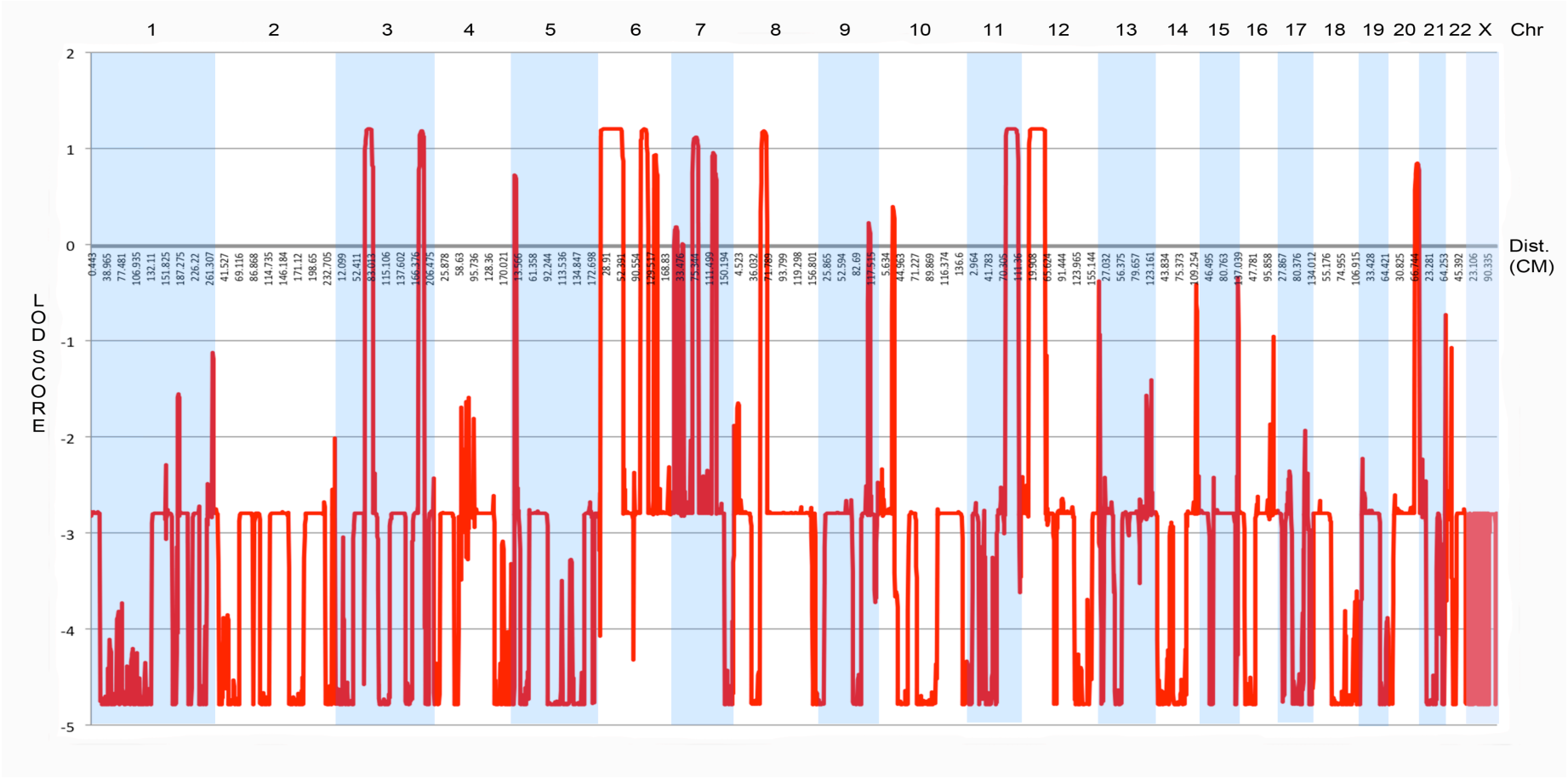

Figure 29 – Plot of genome-wide linkage data. The chromosome numbers are shown across the top of the plot. The central x-axis represents chromosomal distance in centiMorgans (CM). The y axis indicates the LOD score. The maximal LOD score was 1.204 and there are 8 peaks with a LOD score of greater than 1.

8.3.3 CNV Analysis

CNV analysis revealed 8 small CNVs common to all three affected individuals (table 25). Most of these were within intergenic regions and contained no known genes, including the only homozygous deletion. The three shared CNVs that did contain genes, on chromosomes 7, 9 and 16, were heterozygous duplications and thus unlikely to be the cause of this ostensibly autosomal recessive condition.

Table 25 - Copy number variants (CNVs) common to all three affected siblings. 2 is the normal copy number, such that 0 indicates a homozygous deletion, 1 a heterozygous deletion, and 3 a heterozygous duplication. For each CNV, genomic coordinates, size (in kilobases), the number of SNPs the call is based upon and the number of genes in the region are given.

Chr	Start	Stop	Length (Kb)	SNPs	Copy Number	Genes
5	117418263	117421055	2.7	4	1	0
6	121775368	121858807	83.4	16	3	0
7	27127364	27227300	99.9	23	3	14
8	43689385	43793527	104.1	6	1	0
9	107507950	107518947	10.9	4	3	1
11	81181640	81194909	13.2	7	0	0
13	22320753	22437563	116.8	14	3	0
16	1247677	1315264	67.5	10	3	5

8.3.4 Exome Sequencing and Variant Filtration

Exome sequencing generated an average of 81270305 unique reads. Based on the CCDS hg19 definition of the exome, average coverage at 2x, 10x and 20x read depth was 94.3%, 91.2% and 86.7%, respectively, with the mean coverage across the whole exome of 66 reads. This translated to an average variant count of 22877 variants per exome.

In order to isolate potentially pathogenic variants, we applied a systematic filtering procedure to the data (figure 30). We began by selecting only those variants that were present in the exome data of all three affected family members for analysis.

Synonymous variants and variants recorded in dbSNP135 were initially removed. We then filtered out any variant present at a global minor allele frequency of ≥1% in a range of publically available databases of sequence variation (1000 Genomes, Complete Genomic 69 Database and NHLBI Exome Sequencing Project database) as well as those found in 2 or more our own in-house exomes from individuals with unrelated diseases (*n*=200).

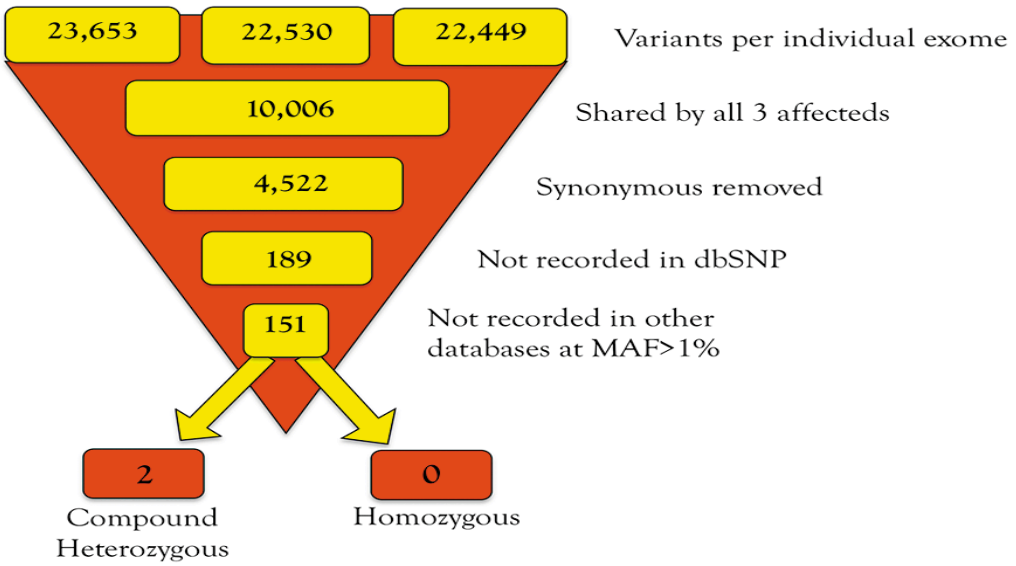

Figure 30 – Graphic illustration of the filtering process applied to the exome datasets in order to select out likely candidate causal variants for this ostensibly autosomal recessive condition. Other databases = 1000 Genomes, NCHLBI Exome Sequencing Project and the Complete Genomic 69 database.

8.3.5 Potentially Causal Candidate Variants

No shared, potentially pathogenic homozygous variants were identified by this method, suggesting that our hypothesis that compound heterozygous mutations were the likely cause of disease was correct.

Two genes with shared compound heterozygous variants were identified: *LRRC16A*, on chromosome 6, which haboured two missense mutations (c.293C>T [p.A98V] and c.437A>G [p.D146G]); and *ATM*, on chromosome 11, which haboured a 4 base-pair, frameshift deletion (c.7886_7890del) and a missense mutation (c.6154G>A [E2052K]).

LRRC16A has not been associated with any disease at present. It encodes a 1,374 amino acid protein that contains 11 leucine-rich repeats and a c-terminal actin capping protein (CAPZA2) binding domain. This CAPZA2-binding domaine appears to be both necessary and sufficient for the protein to inhibit actin capping by CAPZA2 [366]. The gene is predominantly expressed in the kidney, though it is detected in many tissues at low levels, and it appears to be involved in lamellipodia protrusion [366]. Knockout of this gene in mice does not result in any detectable phenotype [367]. Both of the mutations detected in this study are located near beginning of the protein, 99 amino acids before the beginning of the first leucine rich repeat and over 800 amino acids before the beginning of the critical CAPZA2-binding domain. The c.293C>T mutation is predicted to be benign by both SIFT (score=0.32) and PolyPhen (score=0.005), whilst the c.437A>G mutation is predicted to be deleterious by SIFT (score=0.03) and possibly damaging by PolyPhen (score=0.698). Most importantly, segregation analysis by Sanger sequencing demonstrated that the variants in *LRRC16A* were present in the compound heterozygous form not just in the affected individuals but also in one unaffected sibling (II:3 in figure 28), thus making *LRRC16A* an unlikely cause for the DRD.

The second gene in which we identified compound heterozygous mutations was *ATM*, *ATM* is a gene normally associated with ataxia telangiectasia, a condition in which dystonia can be seen. Notably this was the only compound heterozygous mutation that sat within a linkage peak, occurring in a primary peak of interest on chromosome 11. One of the mutations, a frameshift deletion in exon 54 of the gene (c.7886_7890del), is predicted to lead to nonsense mediated decay of the truncated mRNA product and has been detected in our own patients with ataxia telangiectasia (AT). The other, a point mutation in exon 43 (c.6154G>A), results in a substitution of a positively charged lysine in the place of a negatively charged glutamic acid at position 2052. This position lies in a FAT domain of the protein, which is thought to interact with ATM's kinase domain to stabilise the c-terminus. It is predicted to be deleterious by SIFT (score=0.05) and possibly damaging by PolyPhen (score=0.74) and lie in a region of substantial interspecies conservation with a GERP score for 35 eutherian mammals of

3.81 (see figure 31). Most notably, it has previously been detected in patients with AT in the homozygous state by others, and appears to cause skipping of exon 44 [368].

Figure 31 – Diagram demonstrating complete conservation at the base position c.6154 between 18 amniota vertebrates. This base is the first of a codon specifying glutamate (GAA). Note that despite the 3rd base interspecies 'wobble' between an A and a G, the redundancy of the genetic code means there is nonetheless complete interspecies conservation of the amino acid at this position (p.2056). The same is true of the three codons shown adjacent to this.

8.3.5 Clinical Rephenotyping and AFP Measurement

Given the finding of two pathological mutations in *ATM*, we returned to the family to re-evaluate the pedigree and re-examine the index cases. Re-evaluation of the pedigree revealed that one of the unaffected siblings had since had two daughters who are now affected by clinically typical AT. DNA samples were obtained for this branch of the family and the relevant exons of *ATM* sequenced. This revealed that the unaffected sibling and his wife (who may be related to him, though the family is unaware of this) both carry the 4 basepair, frameshift deletion (but not the missense mutation) in the heterozygous state. One of his daughters, who suffers from clinically-typical AT, is homozygous for the same deletion. We were unable to obtain blood from the other affected daughter. Figure 32 shows the updated pedigree with mutational status.

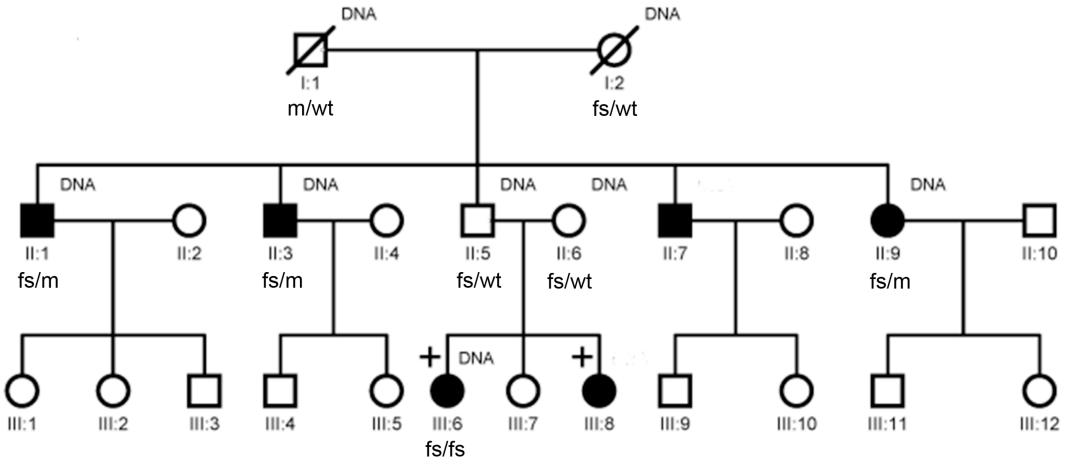

Figure 32 - Updated genetic pedigree in light of the initial exome sequencing results. The individuals marked with a cross (III:6 and III:8) have clinically typical ataxia telangiectasia. Mutational status for ATM is marked below the relevant individuals as follows: wt = wildtype allele; fs = frameshift mutation (c.7886_7890del); and m = missense mutation (c.6154G>A). DNA availability is marked as previously. The marriage between individuals II:5 and II:6 is probably consanguineous.

Re-examination of the three siblings affected by DRD revealed conjunctival telangiectasia (figure 33) but no ataxia or oculomotor signs. They remained responsive to levodopa. In addition, a fourth sibling now exhibited very mild cervical dystonia and dysarthria, but did not wish to participate in the genetic analysis and so his mutational status could not be ascertained.

Figure 33 - Conjunctival telangectasia in a sibling affected by cervical dopa-responsive dystonia.

Alpha fetoprotein is a generally elevated in AT. Therefore, we obtained simultaneous blood samples from all individuals that were willing to donate blood for DNA analysis for estimation of serum alpha fetoprotein levels. Testing revealed that alpha fetoprotein was raised in all three siblings with DRD as well as the daughter of the unaffected sibling who has clinically typical AT (table 26).

Table 26 - Serum alpha fetoprotein measurements for key individuals in the family. Individual labels relate to the genetic pedigree (figure 32). Mutational status is coded as follows: wt= wlidtype allele; fs = frameshift deletion (c.7886_7890del); m = missense mutation (c.6154G>A).

Individual	Alpha Fetoprotein measurement in IU/ml (normal = <7)	Mutational status
II:1	55.42	fs/m
II:3	53.19	fs/m
II:9	97.28	fs/m
II:5	1.23	fs/wt
II:6	1.93	fs/wt
III:6	208	fs/fs

8.4 Discussion

We performed whole exome sequencing and linkage analysis in 3 individuals affected by dopa-responsive dystonia from a Muslim, Indian kindred. We did not find any evidence of significant autozygosity within the family to suggest inbreeding was a factor. Compound heterozygous mutations were identified in two genes – *LRRC16A* and *ATM*. However, the mutations in *LRRC16A* did not lie within a linkage peak and as expected did not thus segregate with disease in the family.

The compound heterozygous mutations in the gene ATM, on the other hand, did lie within a linkage peak of primary interest and were found to be shared by all 3 affected individuals. Sanger sequencing subsequently confirmed that the two mutations were on separate alleles, with each parent carrying only one mutated allele. One mutation, a frameshift deletion in exon 54 of the gene (c.7886_7890del), is predicted to lead to a nonsense mediated decay of the truncated mRNA product and has been observed in individuals with clinically typical AT in our own neurogentics diagnostic laboratory; the other, a point mutation in exon 43 (c.6154G>A), leads to the substitution of a conserved amino acid, is predicted to be deleterious by SIFT and possibly damaging by PolyPhen and has been observed in the homozygous state in patients with clinically typical AT by others [368].

Biallelic mutations in ATM are generally associated with ataxia telangiectasia (AT), an autosomal recessive form of early-onset ataxia associated with oculomotor apraxia, ocular and cutaneous telangiectasia and variable immunodeficiency Patients with classical disease generally develop an ataxic gait in their early childhood and are wheelchair bound by early adolescence. Death often occurs in the second or third decade of life due to malignancies or respiratory failure [369],[370]. However, a variant form of the disease is recognised with either much a later and more slowly progressive course, or, occasionally, where the cardinal features of the disease – ataxia, telangiectasia, and immunodeficiency – are completely absent and the patient presents instead with a movement disorder or peripheral neuropathy [371]. Extrapyramidal disorders appear to dominate the clinical presentation of variant ataxia: in the study by Verhagan et al., 60% had rest tremor, 60% had dystonia and 70% had choreoathetosis.[371].

Notably, a recent publication identified members of three Canadian Mennonite families who presented with early-onset, isolated, predominantly cervical dystonia without frank ataxia or oculomotor apraxia. These individuals harboured a homozygous c.6200C>A missense mutation in exon 43 of ATM [372]. This is the same exon, in which we detected a missense mutation (c. 6154G>A) in our Indian Muslim kindred with dopa-responsive dystonia. Since the frameshift mutation that they carried on the other allele would be expected to lead to nonsense mediated decay of the truncated mRNA product, the allele bearing the c.6154G>A mutation would be the only one expected to reach the point of translation into a protein product. The exact mechanism by which some mutations lead to the variant presentation is not understood. It was originally suggested that, unlike null mutations which leave no functional ATM protein, other mutations – particularly point mutations, which are seen in about 10% – may result only in ATM with a reduced or aberrant functionality. However, other studies have shown that the milder phenotype – including that of pure cervical dystonia mentioned above – can be associated with a very low or even complete lack of detectable ATM activity [372, 373]. It is, however, interesting to note that all individuals with dopa responsive dystonia exhibited a serum alpha-fetoprotein measurement in the intermediate range (approximately 50 -100IU), whereas the individual with clinically typical AT (who habored two frameshift mutations) had a serum alpha fetoprotein level that was four-to-two times this level (208IU).

Furthermore, the relationship between ATM and extrapyramidal features of the disorder remains unclear. Abnormalities of the basal ganglia have been noted on studies involving brain imaging (T2 putaminal hyperintesity on MRI [374]; hyperechogenicity on transcranial ultrasound [372]) or neuropathological examination (Lewy Body formation [375]) carried out in a handful of cases of classical or variant AT. However, the numbers are too small and the changes described often too non-specific to draw any meaningful conclusions. Nonetheless, ATM knockout mice have consistently been shown to exhibit late-onset degeneration of nigro-striatal neurons, suggesting a possible mechanism underlying the development of extrapyramidal features in AT [376].

The identification of dopa-responsive dystonia as a manifestation of AT has double significance. Firstly, it suggests that at least a subset of the dystonic features seen in classical and variant AT may be ameliorated by treatment with L-dopa, possibly leading to an improvement in the patient's functional ability. Secondly, despite the milder phenotype, the rate of malignancy in variant AT appears, nonetheless, to be significantly elevated and patients remain at increased risk from exposure to radiation used diagnostically or as a treatment [371, 377]. Therefore, unexplained dopa-responsive dystonia – particularly cervically located and of young onset joins the list of conditions in which clinicians should consider, at the very least, routine assay to look for raised alpha fetoprotein. In cases of high suspicion, in which unexplained dopa-responisve cervical dystonia is combined with other features of variant AT, such as clumsiness, a sensorimotor neuropathy or early malignancy, chromosomal fragility studies or genetic testing may be warranted.

CHAPTER 9:

Exome Sequencing in an Autosomal Recessive Complex Neurological Disorder Including Bilateral Visual Failure

9. Exome Sequencing in a Family with An Autosomal Recessive Complex Neurological Disorder Including Visual Failure.

9.1 Introduction

Most of the families presented so far have been characterised by a relatively simple phenotype. The clinical picture has generally been dominated by a single neurological sign – dystonia – suggesting dysfunction confined to particular neurological circuits, most likely within the basal ganglia. In some respects, this can be helpful in trying to find new disease genes: a candidate can be prioritised as having higher biological plausibility if its pattern of expression within the brain matches the pattern of regions or structures that are dysfunctional in the disease state. The dystonia gene *ANO3*, detailed in chapter 5, serves as a convenient example here. It is much more highly expressed in the striatum, a structure intimately linked to many movement disorders including dystonia, than in any other region of the brain, suggesting that its main biological function is related to this particular neuroanatomical region. This lends biological plausibility to the disease-gene association and suggests a reason why mutations in this gene can cause dystonia with no evidence of more widespread neurological dysfunction.

Many neurological disorders, however, present with a complex phenotype, indicating dysfunction of multiple systems within the brain or, alternatively, of disparate parts of the nervous system. The list of genetic causes of such complex neurological disorders is long (see section 3.11 for a list of those in which dystonia is a feature) making accurate diagnosis challenging at times. Thus, in order help them arrive at the correct diagnosis, clinicians often pick a particular key symptom or sign on which 'to hang' their list of differential diagnoses. Progressive bilateral visual failure is one sign that can serve as a useful hook. In the context of a patient suspected of having a genetic disorder, bilateral visual failure significantly reduces the list of possible diagnoses and even suggests a likely pathogenetic mechanism. Whether the sole or predominant sign of the disorder, as in autosomal dominant optic atrophy or Leber's hereditary optic atrophy, or whether simply one of numerous signs suggestive of more widespread neurological dysfunction,

progressive bilateral visual failure is highly suggestive of underlying mitochondrial dysfunction. Although mitochondrial DNA is not routinely sequenced as part of current NGS approaches, many of the genes required for normal mitochondrial function are, in fact, autosomally encoded and can thus be examined by whole exome sequencing. With respect to gene discovery, consanguineous kindreds exhibiting signs suggestive of mitochondrial dysfunction, in which the disease is most likely to have resulted from autosomal recessive inheritance, offer a particularly good prospect of success: the power of homozygosity mapping and whole exome sequencing can be further boosted by reasonable *a priori* hypotheses regarding the likely cellular function of the disease gene.

We identified a family in which two of three children were affected by an undiagnosed severe infantile-onset complex neurological phenotype that included bilateral visual failure. The family reported consanguinity, suggesting a probable nuclear-encoded homozygous recessive variant as the underlying genetic cause. We set about trying to identify this causal variant using a combination of whole exome sequencing and homozygosity mapping, based on the hypothesis that the disease gene would most likely be involved in mitochondrial function.

9.2 Subjects, Materials and Methods

9.2.1 Clinical Details of the Index Family

Participants were drawn from a small Pakistani kindred under the care of the Prof. Kailash Bhatia at the National Hospital and Neurosurgery. The pedigree of the family is shown in figure 34. The parents of the affected individuals were first cousins.

Individual II:1 is currently 16 years old. He has been affected by his illness since his infancy and is currently confined to a wheelchair. The most prominent finding on examination is severe action-induced myoclonus, resulting in constant and disabling jerks of all four limbs. Neurophysiological testing has failed to demonstrate a cortical correlate to his myoclonic movements suggesting they are subcortical in origin. On examination of the cranial nerves, eye movements are abnormal with difficulties in

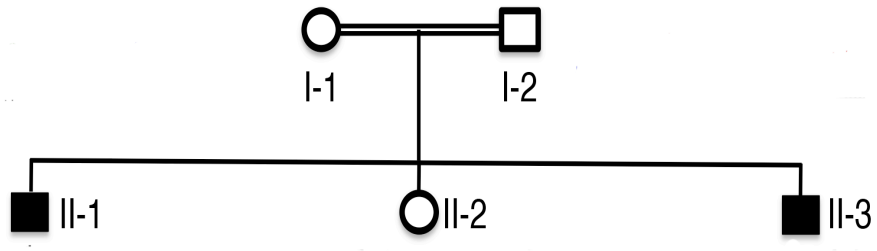

Figure 34 – Core genetic pedigree of the index family. Affected individuals are shown as filled symbols. The marriage between individual I:1 and I:2 is consanguineous and the individuals concerned are first cousins.

initiation of saccades, suggesting an incipient gaze palsy. Vision is impaired bilaterally with evidence of rod-cone dystrophy on the electroretinogram. On examination of the limbs, there is reduced power distally, with striking wasting of the calf muscles bilaterally and clawing of the hands, with accompanying sensory loss in a glove and stocking distribution. On nerve conduction studies and electromyogram have confirmed the clinical impression of an axonal motor-sensory neuropathy. Finally, MRI of the brain has shown cerebellopontine atrophy only.

Individual II:3 also developed signs of neurological illness in his infancy. With time this developed into the same cluster of symptoms, consisting of visual impairment, eye movement abnormalities, action myoclonus and peripheral neuropathy. He is currently 13 years of age and is also wheelchair bound.

Individual II:2 and both parents (individuals I:1 and I:2) are well with no signs of neurological illness.

9.2.2 Whole Exome Sequencing and Generation of Coverage Statistics

DNA was extracted from whole blood samples obtained from all three affected siblings and both parents. DNA from one affected sibling was used to perform whole exome sequencing using Illumina's TruSeq (62Mb) DNA sample preparation and exome enrichment kits as per section 4.15. Reads were subsequently aligned and annotated

using the standard approaches set out in sections 4.16. Coverage statistics for the exome as a whole were obtained using Picard.

Coverage across regions of shared homozygosity was calculated using BedTools against the CCDS definition of the exome and expressed as the percentage of bases target covered by at least one read. Briefly, using Ensembl's Biomart function, a list of every exon of every CCDS gene within each homozygous region was obtained, along with its start position, end position and strand. This information was used to construct a simple BED file of the desired targets within the region. Coverage of these target regions was then checked against the BAM file for the exome using BedTools 'coveragebed' function.

9.2.3 Genotyping and Autozygosity Mapping

Genome-wide genotyping data was obtained for all three affected siblings using the OmniExpress platform, which utilises approximately 500,000 markers, as per section 4.12. Autozygosity mapping was performed as per section 4.13.

9.2.4 Filtration of Variants Detected by Exome Sequencing

In view of the apparently recessive inheritance pattern and history of consanguinity, we initially selected all homozygous variants for consideration. Subsequently, synonymous variants not likely to affect a canonical splice site (i.e. those not within 10 bases, in either direction, of the intron/exon boundary) were discarded. Given the rarity of autosomal recessive dystonia, we further hypothesised that the causal variant was unlikely to be found in any database of normal sequence variation. However, in order to minimize the possibility of incorrectly assigning causality, we filtered out those variants that were recorded in the 1000 Genomes, NHLBI or Complete Genomics 69 datasets at a minor allele frequency of greater than 0.5%. Given that a variant found at even this frequency would be expected to be occur naturally in the homozygous state in around 1 in every 160,000 births, it seemed distinctly unlikely that we would risk filtering out the causal variant with this cut-off. Subsequently, variants that were located in a region of shared homozygosity were selected as potentially causal. No

filtration was performed on the basis of *in silico* predictions of pathogenicity or conservation scores.

9.2.5 Generation of Nucleotide Multispecies Protein Alignments

In order to assess conservation of amino acid sequence, Uniprot, FlyBase and WormBase were interrogated for orthologous protein sequences. Actual alignment of the sequences obtained in this manner was performed using the freely available web-based application ClustalΩ.

9.2.6 Generation of Regional Gene Expression Data

Information on organism wide and brain region specific gene expression was collated as per section 4.19.

9.3 Results

9.3.1 Whole Exome Sequencing

Exome sequencing produced excellent coverage against the target region. 94.6% of target bases were covered at a read depth of 2, 83.3% were covered at read depth of 10 and 81.1% at a read depth of 20; the mean read depth across the exome was 46. In total exome sequencing generated 94,291,782 reads, of which 62,819,270 were successfully aligned to the reference genome. A total of 21,696 variants were detected in the exome of this individual.

9.3.2 Autozygosity Mapping, Coverage of Homozygous Regions and Variant Filtration

Autozygosity mapping demonstrated that runs of homozygosity greater than 1Mb were relatively common in all three siblings, with the largest single run of autosomal homozygosity stretching for 64Mb. This suggested the possibility of extensive consanguinity within the family. Runs of homozygosity shared between both affected siblings but not present in the unaffected sibling are summarized in the table 27 below.

Initially, coverage of the homozygous regions was checked as set out in section 9.2.2. Coverage was defined as the percentage of target bases covered by at least one read.

Coverage was generally excellent, with all regions of interest showing coverage of >80% and all but two showing coverage of >90%. Nonetheless, it should be noted that manual inspection of the uncovered bases revealed that many of these fell within the UTRs of target genes meaning that coverage of coding bases is probably somewhat better than shown below.

Subsequently, the number of variants of all types detected in each homozygous region was determined. Finally, variant filtration was performed as per section 9.2.4 to generate a list of candidate homozygous variants. Those that lay with a run of homozygosity were defined as potentially causal. This information is also summarized as part of table 27.

Table 27 - Genomic position and size of runs of shared homozygosity. The number of CCDS genes in each region, along with percentage coverage on exome sequencing, total variants detected and potentially causal variants remaining after exome sequencing in also shown.

Chr	Start	End	Length (Mb)	CCDS Genes	% Coverage	Variants Detected	Potentially Causal
1	208369825	214588645	6.22	38	90.0%	31	1
2	95395757	98334450	2.94	33	92.7%	37	0
2	226092297	226814991	0.72	1	96.4%	1	0
3	79919894	81187867	1.27	0	-	0	0
3	186961478	190330777	3.37	13	83.0%	12	0
4	88882095	112909820	24.03	86	86.8%	62	0
5	93798912	121463284	27.66	73	90.6%	62	1
8	85802488	87003929	1.20	9	91.3%	4	0
10	125602021	127251058	1.65	9	94.8%	4	0
11	63691049	67526279	3.84	147	92.8%	116	3
11	126303742	134934063	8.63	30	93.7%	15	0
12	50365997	51595505	1.23	20	93.7%	21	0

9.3.3 Overview of Potentially Causal Variants

After appropriate data filtration, five homozygous variants were identified that met the criteria for being potentially causative, on chromosomes 1, 5, and 11. These variants are summarized in the table 28 and discussed individually in the following sections.

Table 28 – Details of potentially causal variants remaining after filtration of exome data, including chromosomal location, sequence change, amino acid substitution, zygosity, novelty, minor allele frequency (if previously observed) and number of reads covering the variant.

Chr	Position	Gene	Change	Zygosity	Novel?	MAF	Read depth
1	212911871	*NSL1*	c.725C>T p.T242I	Hom	Yes	-	17
5	110081998	*SLC25A46*	c.413T>G p.L138R	Hom	Yes	-	15
11	64003467	*VEGFB*	c.286C>G p.Q96E	Hom	No	0.0018	49
11	65658601	*CCDC85B*	c.347G>T p.C116F	Hom	Yes	-	2
11	66392434	*RBM14*	c.1087T>G p.S363A	Hom	No	0.0009	68

NSL1:

NSL1 encodes a protein that forms part of the MIS12 complex, which is required for normal chromosome alignment and segregation [378].

The variant identified by exome sequencing in this family is a C to T substitution at position 725 of the cDNA (ENST00000366977) resulting in the incorporation of an isoleucine in place of an threonine at position 242 in the protein product. The affected base is well conserved (PhyloP = 3.759; PhastCons = 0.99). The variant is not located in a key domain of the protein according to Uniprot, but the affected threonine is a target for phosphorylation [379]. It is predicted to be tolerated by SIFT (0.111), deleterious by Provean (-3.362), probably damaging by PolyPhen2 (0.997), and disease-causing by MutationTaster (0.999).

As might be expected for a gene involved in chromosome alignment, publically available systemic expression data for *NSL1* based on EST tags shows that it is widely expressed throughout the body, including the brain (figure 35, left side). Our own in house region-specifc brain expression data demonstrate relatively good expression in all areas of the brain with no particular difference between different structures (figure 35, right side).

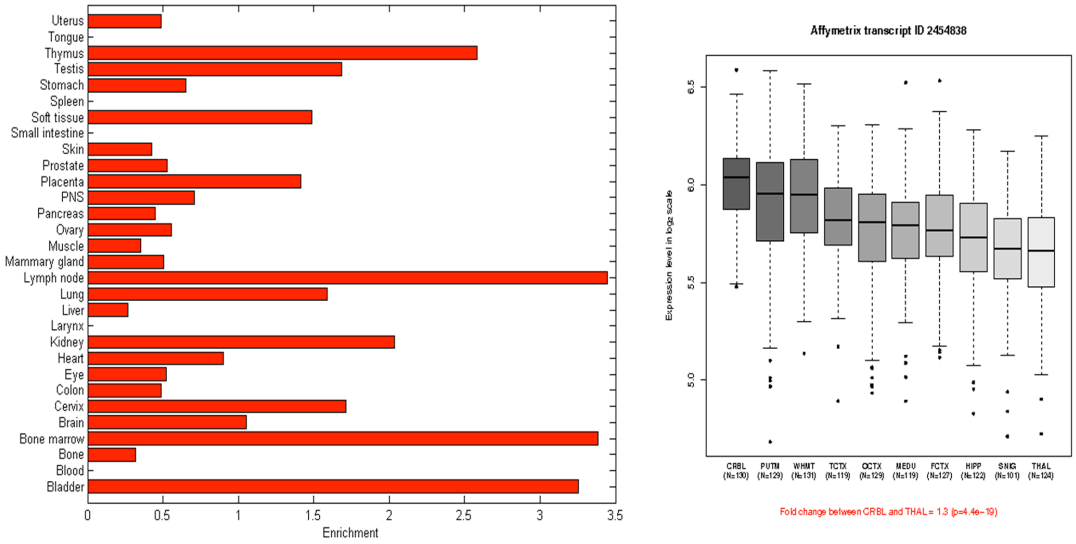

Figure 35 - Expression data for the gene *NSL* in man. The panel on the left shows the organism wide expression data based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 9.2.6, across the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

SLC25A46:

SLC25A46 encodes a protein that belongs to the SLC25 family of mitochondrial carrier (MC) proteins. These proteins are embedded in the inner mitochondrial membrane and catalyse the transportation of a variety of substances. In doing so, they provide a vital link between that mitochrondria and cytosol.

Though MCs display a great diversity of transported solutes (i.e., a large variety of metabolites, nucleotides and coenzymes), all SLC25 members share common sequence features: a tripartite structure, a 3-fold repeated signature motif, and six transmembrane α-helices (two in each of the three repeats; as shown in figure 36) [380].

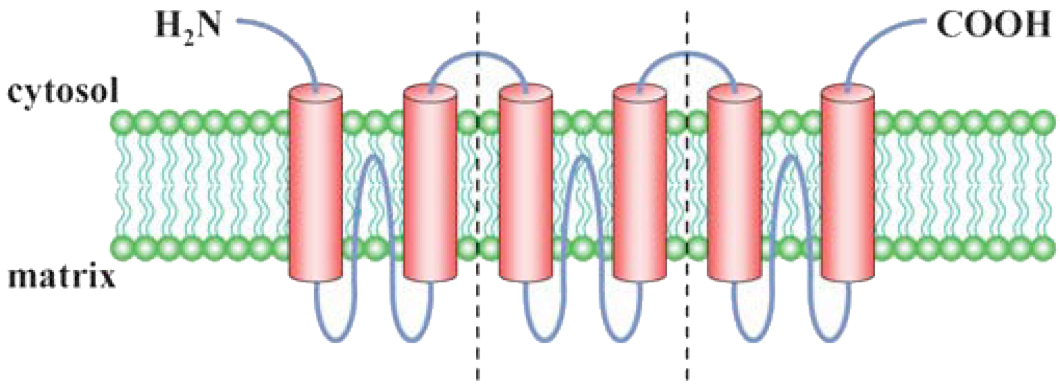

Figure 36 - Topological model of the mitochondrial carrier proteins. Six helices traverse the inner mitochondrial membrane with the C- and N-termini facing the cytosol. The whole sequence is divided into three domains, each with two transmembrane helices connected by a long hydrophobic matrix loop that is assumed to protrude into the membrane. Reproduced from Palmieri (2004) [381].

SLC25A46, along with *SLC25A38* and *SLCA25A44*, is one of a small group of MCs that are quite different from other members of the family and as well as from each other [382]. Phylogenetically, it is most closely related to *SLC25A1*, which is also known as Citrate Transporter Protein on account of its own preferred substrate (see figure 37). However, identity between even these two members of the family is only 13%. The substrate of *SLC25A46* is unknown at present.

The importance of this family of proteins to proper cell functioning is attested by the fact that mutation of many members of this gene family is associated with disease (summarised in table 29). Moreover, as might be expected with mutations affecting a mitochondrial pathway, the resultant disease often affects that nervous system.

Table 29 - Current known disease associations for various members of the *SLC25A* family of mitochondrial carrier proteins. Note that many diseases affect the nervous system.

Gene	Substrate	Disease
SLC25A3	Phosphate	Mitochondrial phosphate carrier deficiency [383]
SLC25A4	ADP/ATP	AAC1 deficiency [384]
		Autosomal dominant progressive external ophthalmoplegia. [384]
SLC25A12	Aspartate/glutamate	AGC1 deficiency with associated global hypomyelination [385]
SLC25A13	Aspartate/glutamate	AGC2 deficiency (neonatal intrahepatic cholestasis/type II citrullaemia) [386, 387]
SLC25A15	Ornithine/citrulline	Hyperornithinaemia-hyperammonaemia-homocitrullinuria syndrome [388]
SLC25A19	Thiamine pyrophosphate	Congenital Amish microcephaly [389]
		Neuropathy with striatal necrosis [390]
SLC25A20	Carnitine/acylcarnitine	CAC deficiency [391]
SLC25A22	Glutamate	Early epileptic encephalopathy III [392]
SLC25A38	?Glycine/alanine	Pyridoxine-refractory autosomal recessive sideroblastic anaemia [393]

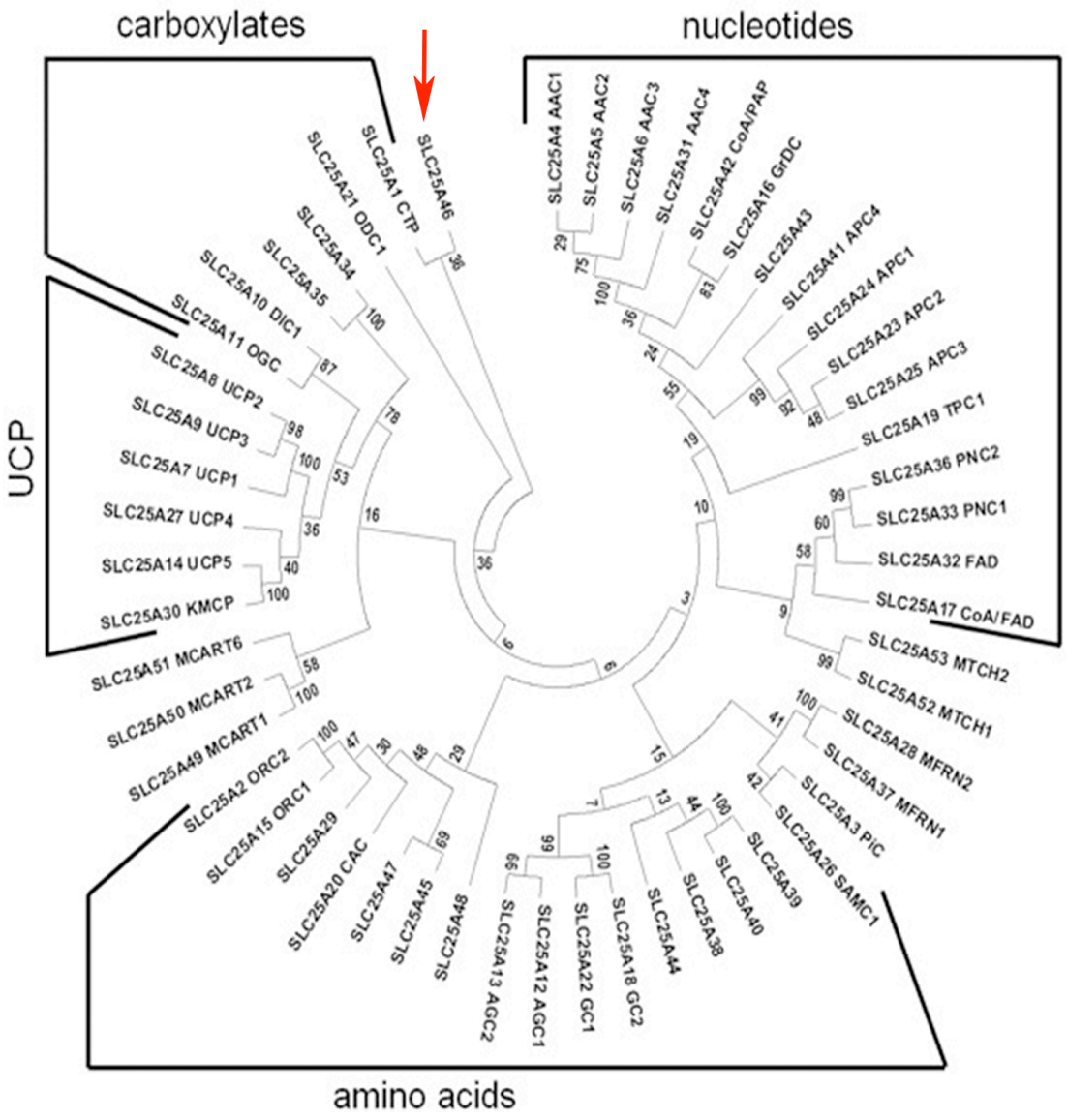

Figure 37 - Phylogenetic tree of human MCs. The tree was originated from ClustalΩ multiple-sequence alignments by using the neighbor-joining method implemented in MEGA5. All 53 carriers of *Homo sapiens* are shown. Bootstrap values for 1000 replicates are reported on each node; gene names and aliases describing carrier function are reported on each leaf node. The clades of uncoupling proteins (UCP) and carriers transporting nucleotides, carboxylates and amino acids are highlighted. This figure and figure legend are reproduced from Palmieri *et al.* (2013) [380]. *SLC25A46* has been indicated by the means of a red arrow.

The variant identified by exome sequencing was situated in the largest stretch of homozygosity in this family. It is a T to G substitution at position 413 of the cDNA (ENST00000366977) resulting in the incorporation of an arginine (large and strongly hydrophilic) in place of a leucine (small and hydrophobic) at position 138 in the protein product. The substitution of a hydrophilic amino acid at this position is likely to be of function relevance since, according to Uniprot, the variant is located in the first of two key SOLCAR (solute carrier) domains of the protein, specifically in the hydrophobic matrix loop which is presumed to imbed in the mitochondrial membrane.

The affected base is well conserved (PhyloP = 4.417; PhastCons = 1) and the affected amino acid is identical in all species down to *C. elegans* (figure 38). It is predicted to be damaging by SIFT (0.001), deleterious by Provean (-4.314), probably damaging by PolyPhen2 (0.972), and a disease-causing by MutationTaster (0.999).

Figure 38 - Multiple sequence alignments for the region of *SLC25A46* harbouring the potentially causative p.L138R variant detected by exome sequencing. The affected base is conserved in all species down to C. elegans. Colours indicate physiochemical properties of amino acids (red = small/hydrophobic; blue = acidic; magenta = basic; green = hydroxyl/sulfhydrl/amine/glycine).

Publically available organism-wide expression data for *SLC25A46* based on EST tags shows that it is widely expressed throughout the body, including the brain (figure 39, left side). Our own in house expression data based on Affymetrix Exon Array profiling of ~130 control brains across multiple areas demonstrates high levels of expression in all areas of the brain (figure 39, right side).

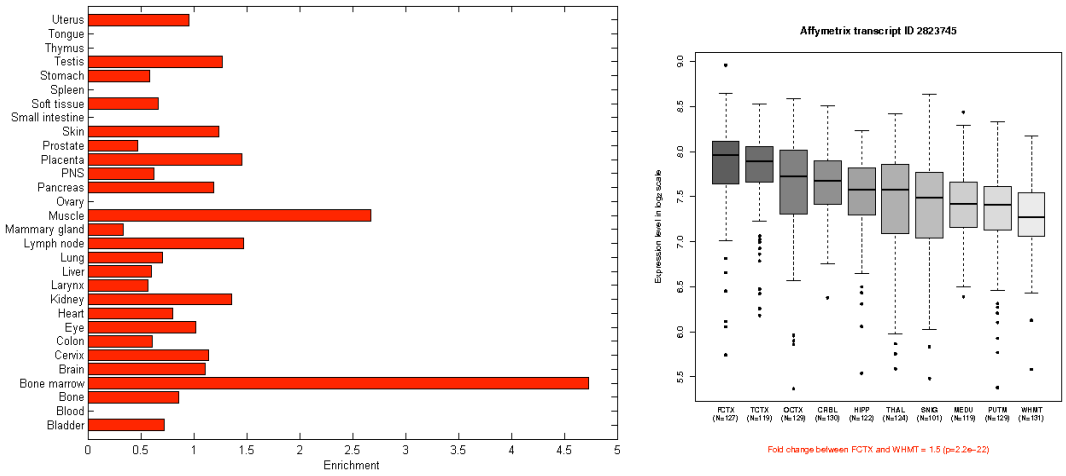

Figure 39 - Expression data for the gene *SLC25A46* in man. The panel on the left shows the organism wide gene expression based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 9.2.6, for the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

In addition, relative expression for this protein has been determined throughout the brain and body of the rat for its ortholog *Slc25a46* [382]. Theses data demonstrate that, although *Scl25a46* is expressed in most tissues throughout the body of the rat, it expression level is much higher in the CNS (see figure 40). Within the rat brain, it is expressed at particularly high levels in the hindbrain (cerebellum, pons and medulla; marked by a red arrow in figure 40), which is interesting given that the affected individuals in this family demonstrated cerebellopontine atrophy on MRI.

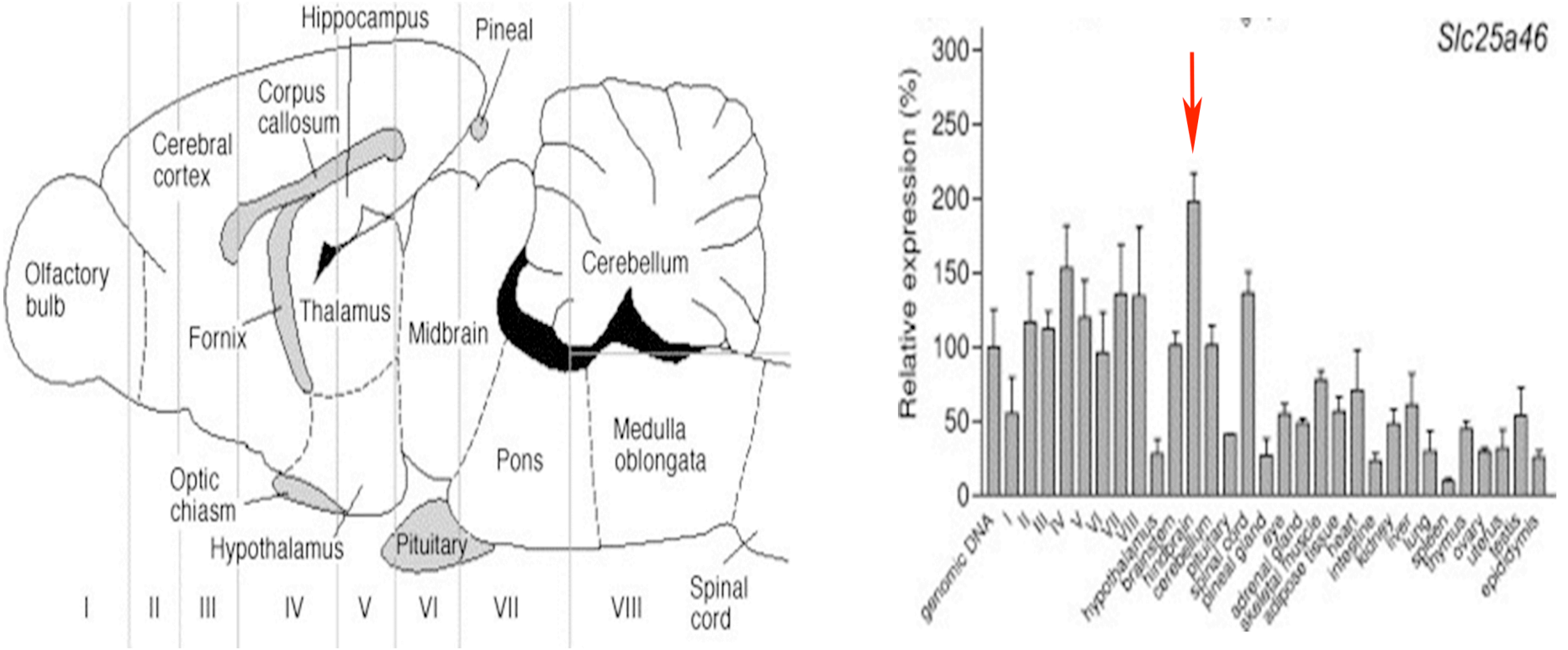

Figure 40 – Expression of the *Scl25a46* in the rat as determined by Haitina *et al.* (2006) [382]. On the left a schematic representation of rat brain coronal sections used for the expression analysis is shown. The sections are marked with Roman numerals. On the right, the relative expression of Scl25a46 (normalised against genomic DNA) is shown for each region of the brain and other body parts by means of a bar graph. Note the particularly high expression of this gene in the rat hindbrain (marked with a red arrow), a region of the brain demonstrated atrophy on the MRI scans of affected individuals in the index family.

VEGFB:

This gene encodes a member of the PDGF (platelet-derived growth factor)/VEGF (vascular endothelial growth factor) family. The VEGF family members regulate the formation of blood vessels and are involved in endothelial cell physiology [394, 395]. This member is a ligand for VEGFR-1 (vascular endothelial growth factor receptor 1) and NRP-1 (neuropilin-1).

From a neurological point of view, VEGFB stimulates neurogenesis in the adult brain. BrdU labeling of immature neurons in the subventricular and subgranular zone is reduced in VEGF-B knockout mice and enhanced in VEGFB-treated neuronal cultures and in rats given VEGFB via ICV injection [396]. VEGFB is also a survival factor and a potent apoptosis inhibitor for neurons. It is capable of protecting retinal and motor neurons from degeneration *in vivo* [397, 398]. It is also expressed in sensory neurons, but, under normal conditions, is not essential for their survival. However, DRG neurons isolated from adult *VEGFB* knockout mice exhibit increased neuronal stress under baseline culture conditions and are also more sensitive to stress stimuli [399].

The variant identified by exome sequencing in this family is a C to G substitution at position 286 of the cDNA (ENST00000309422) results in the incorporation of a glutamic acid in place of a glutamate at position 96 in the protein product. It is not a novel change, being recorded in dbSNP as rs111555072, but its MAF is sufficiently low (0.002) that homozygosity would only be expected to occur in around 1 in a 1,000,000 births. The affected base shows mixed conservation scores (PhyloP = 0.919; PhastCons = 1). This portion of the protein has no homologue in the chicken, frog, drosophila, or c. elegans, and the amino acid is not conserved in the pufferfish or zebra fish, where a threonine is present, or in the duckbill platypus, where a proline is present instead. It is predicted to be damaging by SIFT (0.029), neutral by Provean (-0.304), possibly damaging by PolyPhen2 (0.46), and a polymorphism by MutationTaster (0.999).

Publically available organism-wide expression data for *VEGFB* based on EST tags shows that it is widely expressed throughout the body, including the brain, with particular enrichment in the tongue (figure 41, left side). Our own in-house expression data

based on Affymetrix Exon Array profiling of ~130 control brains across multiple areas demonstrates high levels of expression in all areas of the brain, with the cerebellar being two-fold lower than most other areas (figure 41, right side).

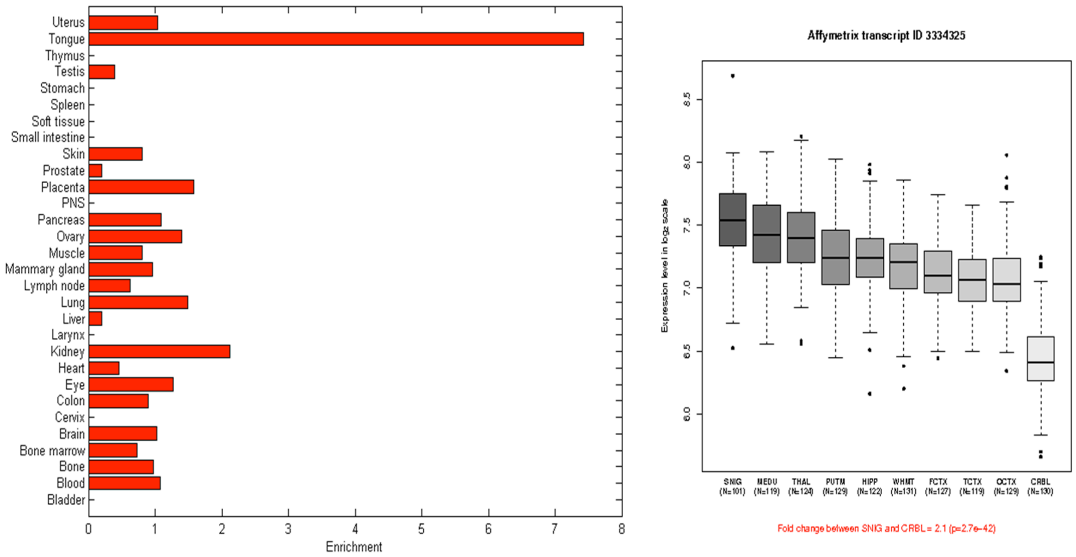

Figure 41 - Expression data for the gene *VEGFB* in man. The panel on the left shows the organism wide gene expression based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 9.2.6, for the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

CCDC85B:

CCDC85B, which stands for 'coiled-coil domain containing 85B', is also known as 'hepatitis delta antigen-interacting protein A' owing to its homology the hepatitis delta viral antigen (HDAg). The product of this gene appears to function as a transcriptional repressor and may inhibit the activity of CTNNB1 in TP53-dependent manner, thus regulating cell growth [400]. It has also been implicated in adipocyte differentiation [401].

The variant identified by exome sequencing in this family is a G to T substitution at position 347 of the cDNA (ENST00000321579) results in the incorporation of a phenylalanine in place of a cysteine at position 116 of the protein product. The affected base shows relatively good conservation (PhyloP = 2.392; PhastCons = 1). The affected amino acid, however, is less well preserved: this portion of the protein has no homologue in the chicken or cat, and the affected amino acid is not conserved in drosophila, frog, pufferfish, zebra fish or *C. elegans*, where a tyrosine is present instead. It is predicted to be tolerated by SIFT (0.079), neutral by Provean (3.138), possibly damaging by PolyPhen2 (0.46), and a disease-causing by MutationTaster (0.999).

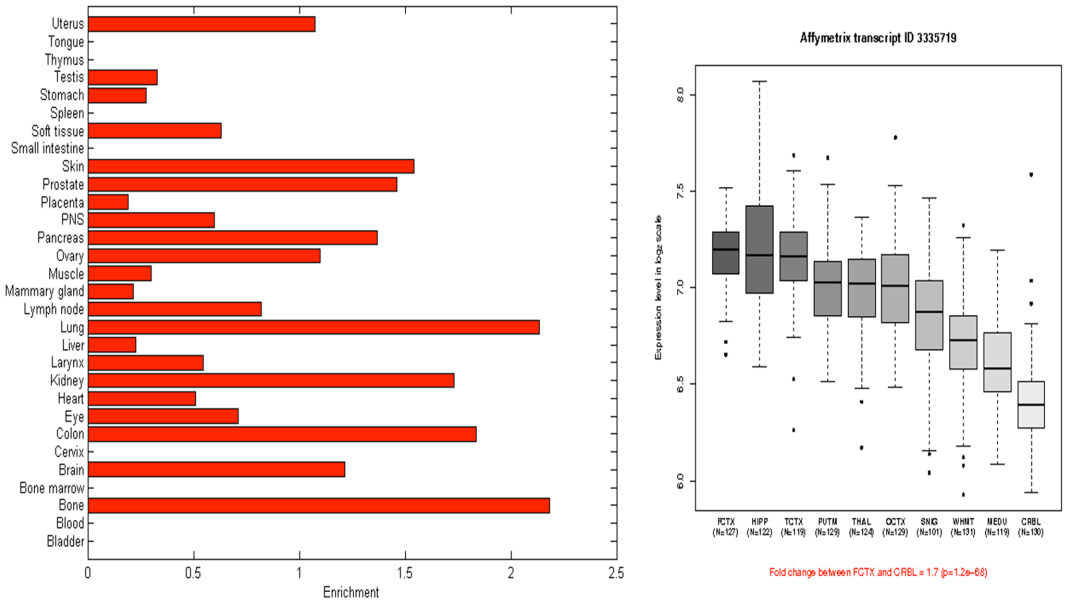

Figure 42 - Expression data for the gene *CCDC85B* in man. The panel on the left shows the organism wide gene expression based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 9.2.6, for the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

Publically available systemic expression data for *CCDC85B* based on EST tags shows that it is widely expressed throughout the body, including the brain (figure 42, left side). Our own in-house expression data based on Affymetrix Exon Array profiling of ~130 control brains across multiple areas demonstrates high levels of expression in all areas of the brain, with particularly high expression in the cortex, hippocampus and deep grey matter (figure 42, right side).

RBM14:

RBM14, which stands for 'RNA binding motif 14', encodes a ribonucleoprotein. Isoform 1 may function as a nuclear receptor coactivator, enhancing transcription through other coactivators such as NCOA6 and CITED1. Isoform 2, functions as a transcriptional repressor, modulating transcriptional activities of coactivators, including isoform 1, NCOA6 and CITED1.

The variant identified by exome sequencing in this family is a T to G substitution at position 1087 of the cDNA (ENST00000310137) results in the incorporation of an alanine in place of a serine at position 363 of the protein product. The mutation is not novel. It is recorded in dbSNP as rs148119991 and has a MAF of 0.0009. Homozygosity would thus be expected for approximately 1 in every 5 million births.

The affected base shows relatively neutral conservation scores (PhyloP = 1.25; PhastCons = 1). The affected amino acid, however, is not conserved in several species and, perhaps most importantly, multiple sequence alignment demonstrates that this amino acid is an alanine (i.e. the same as the mutation found in this family) in our closest ancestor, the chimp, as well as the chicken and the puffer fish. It is predicted to be damaging by SIFT (0.018), neutral by Provean (0.370), possibly damaging by PolyPhen2 (0.46), and a polymorphism by MutationTaster (0.987).

Publically available organism-wide expression data for *RBM14* based on EST tags shows that the gene is widely expressed throughout the body, including the brain (figure 43, left side). Our own in-house expression data based on Affymetrix Exon Array profiling

of ~130 control brains across multiple areas demonstrates high levels of expression in all areas of the brain, particularly the cerebellum (figure 43, right side).

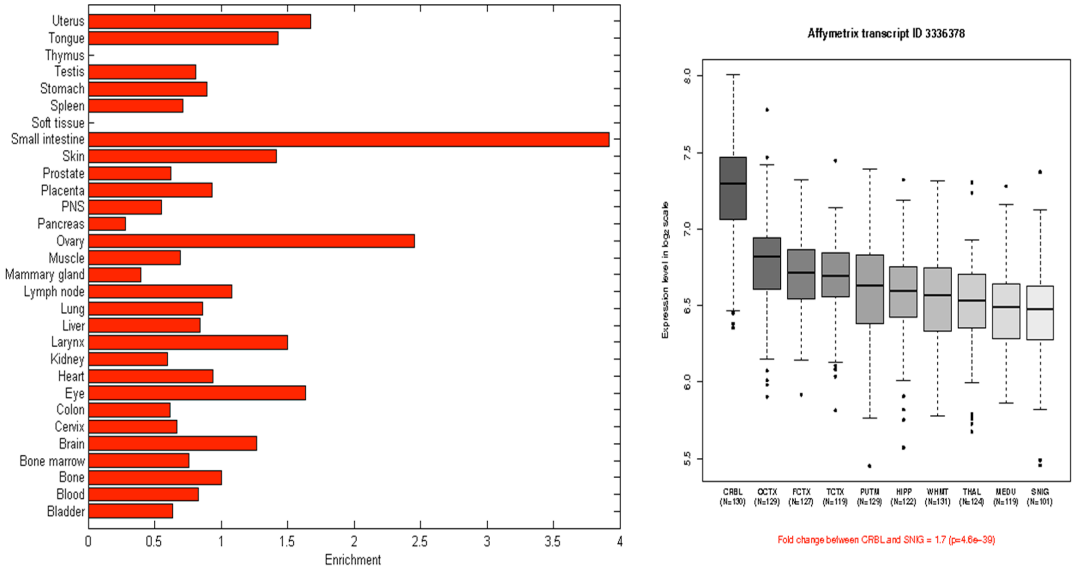

Figure 43 - Expression data for the gene *RBM14* in man. The panel on the left shows the organism wide gene expression based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 9.2.6, for the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

9.3.4 Selection of Candidate Variant for Sequencing

Based on all available information that we were able to curate, it was not possibly to definitively exclude any variant. However, *SLC25A46* stands out as a top hit for several reasons, including:

1) *SLC25A46* encodes a mitochondrial solute carrier and thus fits in with our initial hypothesis that the disorder seen in this family was likely to result from mitochondrial dysfunction, given the occurrence of bilateral visual failure.

2) Impaired mitochondrial function would be expected to produce widespread neurological dysfunction and this is further supported by the frequent occurrence of neurological disease when other members of the *SLC25* gene family are mutated (see table 29). Expression profiling suggests high expression in the CNS and the rat ortholog is particularly highly expressed in the hindbrain, which is, at the very least, interesting in view of the cerebellopontine atrophy seen on MRI in this family.

3) The variant detected in *SLC25A46* is novel, as might be expected if it were the cause of this highly unusual and complicated phenotype, which combines severe myoclonus with visual failure, an axonal motor-sensory peripheral neuropathy and cerebellopontine atrophy. If the condition were caused by one of the non-novel variants (in *VEGFB* or *RBM14*), it is to be expected, based on the MAF of the variants and the current population figures for the UK, that there would be to be approximately 60 similar cases for *VEGFB* and 12 for *RBM14* in the UK alone. Many of which would be seen at quaternary paediatric institutions like Great Ormond Street so that the phenotype might be expected to be more recognizable than it was to the neuropaediatricians.

4) The variant in *SLC25A46* was the only variant to be unanimously predicted to be damaging by all four *in silico* prediction programs (notably with near maximal scores for the certainty of the prediction).

5) The affected amino acid is conserved in all species down to phylogenetically distant creatures such as *C. elegans*, suggesting it may play an important role within the protein. This was not true of any other variant.

6) Finally, the mutation affects the key SOLCAR domain of the protein and results in the incorporation of a large, charged, hydrophilic arginine where a small, neutral, hydrophobic leucine is normally found. The difference in hydrophobicity is particularly relevant given the mutation is located in the first hydrophobic matrix loop and may thus impair its insertion into the membrane.

We therefore decided to take *SLC25A46* forward for screening in a cohort of individuals with a mitochondrial phenotype. As an initial first step, we sequenced exon 4 of *SLC25A46* by Sanger methodology in all members of the index family to ensure

that it segregated as expected on the basis of the homozygosity mapping. The results (shown in figure 44 below) confirmed that the T to G variant was present in the homozygous state only in the two affected individuals and was present in a heterozygous state in all other unaffected members of the family.

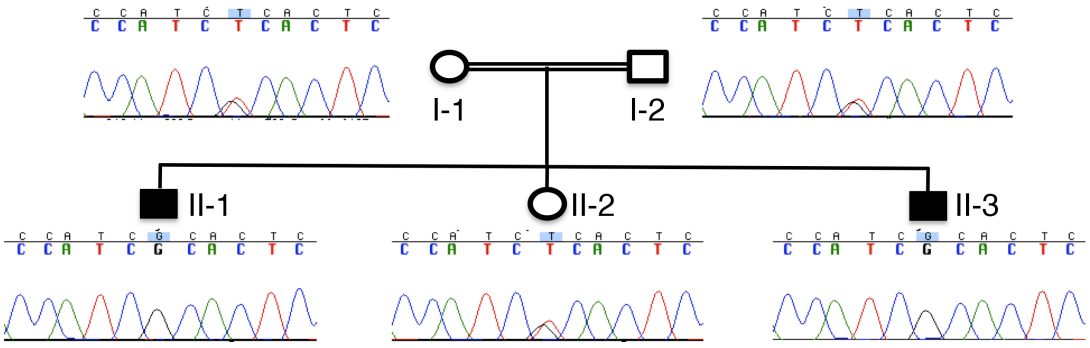

Figure 44 - Segregation analysis in the index family for the homozygous p.L138R variant in *SLC25A46*. As expected the variant is only present in the homozygous state in the affected individuals. The unaffected sibling and both parents are heterozygous for this change.

9.3.5 Selection of Cases for Screening and Sequencing Strategy

SLC25A46 is encoded by 8 exons. The alignment of these 8 exons against a schematic of the protein structure, showing the two key SOLCAR domains, is shown in figure 45. The mutation that we detected in the index family was located in the 4th exon, which encodes part of the first SOLCAR domain. In order to minimise sequencing costs while hopefully maximising our chances of finding a second mutation, we decided to try and focus our efforts on these SOLCAR domains and on exon 4 in particular. Primers were thus designed for exons 2, 3, 4, 5 and 8.

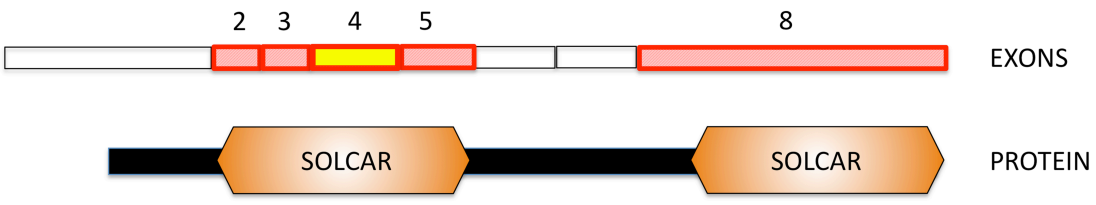

Figure 45 – Schematic representation of the alignment of exons chosen for sequencing against protein structure for *SLC25A46*. The homozygous variant detected by exome sequencing was located in exon 4 of the gene.

We selected 288 cases, with the aim of sequencing exon 4 in all 288 and exons 2, 3, 4, 5 and 8 (which encode the remainder of the two SOLCAR domains) in 192 of these. Cases were selected from an in-house library of research samples on the basis that they had a clinical presentation compatible with a mitochondrial disorder with onset in the first decade of life. Many of these individuals would have been tested for common mitochondrial mutations and deletions, but we were not able to access these results and so it is probable that a proportion will have known mutations. Moreover, we were not able to find any individuals with a clinical description that matched this phenotype of the index family – in particular no case could be found where the clinical details indicated bilateral visual failure in the setting of a more complex neurological disorder. We did not select individuals with pure visual failure and no other neurological signs as we felt this was too far removed from the severe, complicated phenotype seen in the index family. Nonetheless, we attempted to increase our chances by selecting DNA samples if the associated clinical description made any mention of features of the clinical presentation, such as myoclonus, cerebellopontine atrophy or a neuropathy. All in all, however, this only amounted to a handful of cases and the remainder were made up of a mix of disparate presentations compatible with a mitochondrial disorder.

9.3.6 Results of Sequencing of Cohort

We identified only two missense mutations, both in the heterozygous state. One, a G to T substitution in exon 8 of the gene (c.1137G>T) resulting in the incorporation of an aspartate in place of a glutamate at position 379 of the protein, has been previously annotated (rs79149180 in dbSNP) with a minor allele frequency of 0.04. It has been observed in the homozygous state in approximately 1 in 200 normal individuals and thus is unlikely be disease-causing.

The second mutation was novel. It was found in exon 4 (c.446G>T) and affected an amino acid 9 places downstream of the index family's mutation, still within the same SOLCAR domain. It results in the incorporation of an isoleucine in place of a serine at position 149 of the protein product. Multiple sequence alignments for this portion of the protein were generated from orthologous sequences obtained from Uniprot,

FlyBase and WormBase, using ClustalΩ. As is shown in figure 46, conservation is less good at this amino acid position than in the case of the index family's nearby mutation: the zebra fish and C. *elegans* have asparagine (though this has similar physical properties to serine), whilst drosophila has a histidine (which is quite distinct from serine).

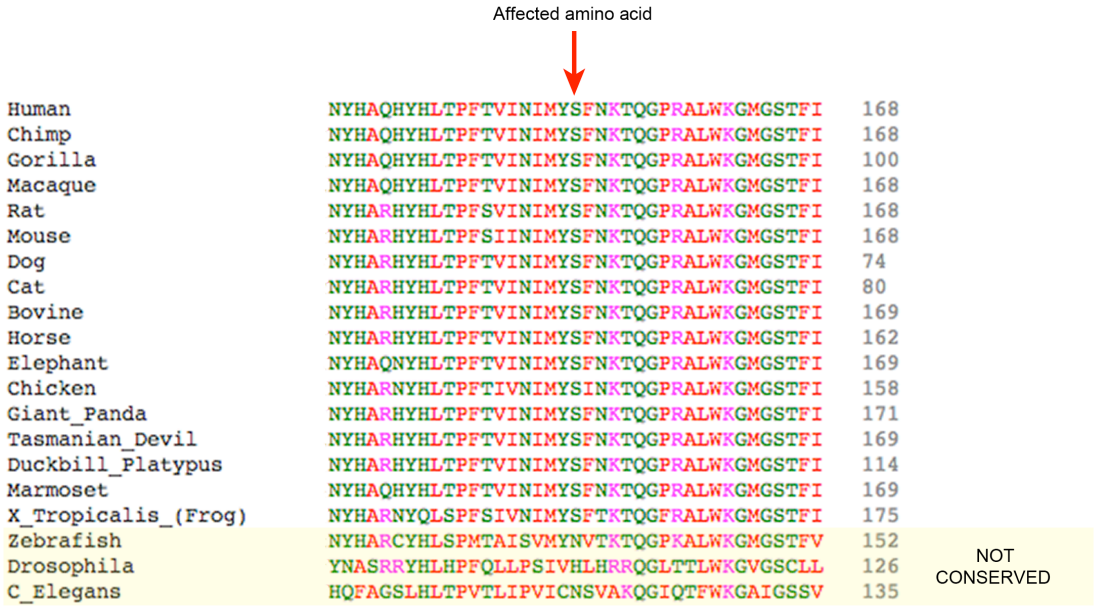

Figure 46 - Multiple sequence alignments for the region of *SLC25A46* harbouring the novel heterozygous p.S149I variant detected by Sanger Sequencing in an individual with a mitochondrial phenotype. The affected base (marked by a red arrown) is conserved in most species, but not the zebrafish, drosophila or the nematode worm, C. *elegans*. Colours indicate physiochemical properties of amino acids (red = small/hydrophobic; blue = acidic; magenta = basic; green = hydroxyl/sulfhydrl/amine/glycine).

In silico predictions of pathogenicity are divided down the middle: the variant is predicted to be tolerated by SIFT (0.14), deleterious by Provean (-2.751), benign by PolyPhen2 (0.092) and disease-causing by MutationTaster (0.997).

Given the presence of a novel heterozygous mutation in the same exon of the gene as our index family, primers were designed for exons 1, 6, and 7 of *SLC25A46* in order to sequence the remainder of the gene in this individual, looking for a second compound

heterozygous change. However, we did not did detect any further potentially causative variants.

9.4 Discussion

In the work presented in this chapter, we set out to elucidate the cause of disease in a consanguineous kindred exhibiting a complex neurological phenotype, which included bilateral visual failure. On the basis of this particular feature, we hypothesised that the causal variant would be likely to be involved in mitochondrial processes as most genetic causes of bilateral visual failure known to date have been linked to mitochondrial dysfunction. Using homozygosity mapping and exome sequencing, followed by appropriate variant filtration based on the presumed mode of inheritance and disease prevalence, we identified 5 possible candidate causal variants. Of the 5 genes affected, only one was involved in mitochondrial function: *SLC25A46*, a mitochondrial carrier protein of unknown substrate. The variant in this gene was completely novel, was universally predicted to be damaging with near maximal probability scores by all 4 *in silico* prediction programs, was situated in the largest stretch of shared homozygosity, and and affected an amino acid that shows absolute conservation across species all the way down to the nematode worm, *C. elegans*. The substitution causes the insertion of arginine, which is the second largest (molecular weight 174.2 Daltons), most highly charged (pKa 12.48), and most hydrophilic (hydrophobicity† -4.5) of all the amino acids, in the place of leucine, which is relatively small (molecular weight 131.18 Daltons), neutral, and strongly hydrophobic (hydrophobicity +3.8) leucine. This difference in hydrophobicity, in particular, may be of relevance as the affected amino acid sits in a solute carrier domain of the protein – more precisely, at the very centre of the hydrophobic matrix loop that is presumed to loop back and imbed into the mitochondrial membrane (see figure 36). The presence of the charged and strongly hydrophilic arginine may, hypothetically, impair the ability of the matrix loop to insert into the membrane, disrupting transporter structure and function. The gene is well expressed in the central nervous system and, in the rat at least, is particularly highly

† Hydrophobicity of amino acids is represented as described in Kyte and Doolittle (1982) [402]. Positive values indicate hydrophobicity; negative values indicate hydrophilic amino acids. The scale ranges from +4.5 (isoleucine) to -4.5 (arginine).

expressed in the hindbrain, an area that showed focal atrophy on MRI scans of affected individuals in our family. Although *SLC25A46* has not yet been associated with disease, several members of the *SLC25* gene family have been and neurological disease is a frequent consequence. Thus, the secondary information collated on this gene and the variant detected therein suggest a high level of biological plausibility for causality with respect to the disease seen in the index family.

Unfortunately, it has not been possible to move forward from this point for two reasons. Firstly, the disease phenotype is very rare and thus we were unable to find any samples in our research database that were labelled as having suffered from similar symptoms. In particular, no sample could be found where the clinical description explicitly stated that there was bilateral visual failure in the context of a more complicated neurological phenotype. In part, this may have resulted from an issue with labelling – most samples with suspected mitochondrial disease are labelled simply as 'mitochondrial disease' without any further elaboration. In the absence of any samples from individuals with a good match for index phenotype, we were forced to select samples that were labelled as mitochondrial disease where disease onset had been in infancy. Any sample that mentioned mitochondrial disease as well as neuropathy or myoclonus was included. With such a poorly matched cohort in the context of such a rare disease, it is not particularly surprising, therefore, that we did not identify a second independent kindred with biallelic mutations in this gene. Secondly, at the outset of this research, the family had made clear that, although they were happy to participate in further genetic studies, they did not wish to undergo any further invasive procedures. They felt, understandably, that merely knowing the name of the gene responsible for the disease in their family was not worth the distress and inconvenience of any further invasive procedures since the chance of this knowledge leading to any sort of novel treatment in the near future that might directly improve their children's condition was minimal. This essentially ruled out the possibility of functional studies in patient cell lines (derived either from muscle or fibroblasts) that might have otherwise been able to demonstrate that mitochondrial dysfunction was indeed the underlying cellular phenotype.

In summary, the role of mutations in *SLC25A46* as the cause of neurological disease in this family remains unproven. The function of the gene and character and position of the variant detected therein strongly suggests – to me, at the very least – that it probably is the causal gene and, at the time of writing, attempts are being made to assemble a cohort for screening that is better matched to the index families phenotype by collaborating with the researchers of the neuromuscular unit at the Institute of Neurology. It is my understanding that they may have a cohort of individuals with a mitochondrial phenotype characterised by young onset rod-cone dystrophy in whom mutational screening of all known genes has been negative.

CHAPTER 10:

Exome Sequencing in Generalised Dystonia with Bilateral Striatal Necrosis

10. Exome Sequencing in Generalised Dystonia with Bilateral Striatal Necrosis

10.1 Introduction

Bilateral striatal necrosis, characterised by symmetrical degeneration of the caudate and/or putamen, can result from a diverse range of conditions, including toxic, infectious, metabolic and neurodegenerative disorders [403]. Clinical features reflect the intimate involvement of these structures in movement and cognition: choreoathetosis, dystonia, spasticity, dysarthria, dysphagia and cognitive decline are all recognised manifestations. In familial cases, mutations in five genes have so far been linked to this condition: *ATP6* [404], *ND6* [405, 406], *NDUFV1* [407], *NUP62* [408], SLC25A19 [409]. The genes *ATP6*, *ND6* and *NDUFV1* are all involved in the harnessing of chemical energy by the mitochrondrial respiratory chain; *SLC25A19* encodes a mitochondrial thiamine pyrophosphate carrier; and *NUP62* is an essential component of the nuclear pore complex, which gates the flow of macromolecules between the nucleus and cytoplasm.

We identified an English kindred in which two of the four siblings were affected by familial young onset generalised dystonia. Previous MRI scans of the brains of both affected individuals had demonstrated symmetrical putaminal hyperinstenites representative of limited bilateral striatal necrosis. The remaining two siblings and both parents were free of any neurological symptoms, suggesting probable autosomal recessive inheritance. Despite extensive investigation, including sequencing of genes known to cause both isolated and complex forms of dystonia, the cause of their disease had not been identified. The family accepted the opportunity to participate in this research using exome sequencing and linkage analysis in order to attempt to establish the nature of the underlying genetic lesion

10.2 Subjects, Materials and Methods

10.2.1 Clinical Details of the Index Family

Participants were drawn from a small English kindred. The genetic pedigree of the core family is shown in figure 47. There was no report of any similar condition either in the

extended family. The family did not report any consanguinity and both parents had originated from separate parts of the country.

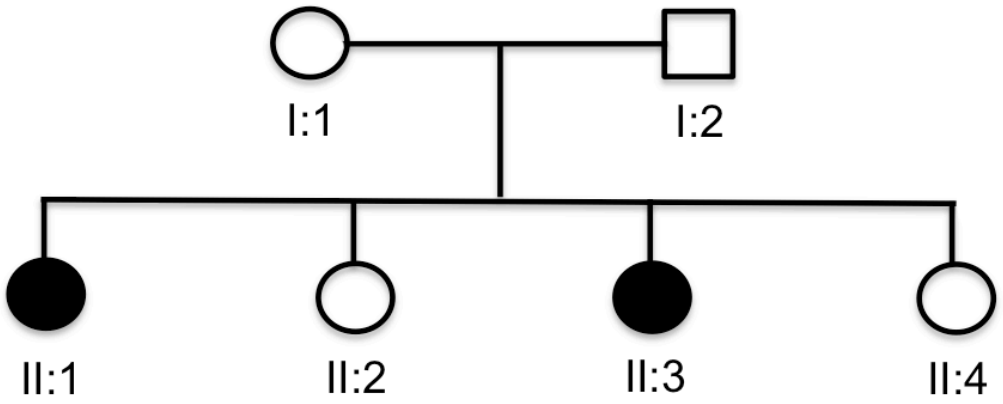

Figure 47 - Abbreviated genetic pedigree of the index family. All siblings were in their late teens or early twenties and did not have children. There was no report of any individual with a similar neurological disease in the extended family. The family did not report any consanguinity and the parents originated from geographically distant regions of the country.

Individual II:1 was born by normal vaginal delivery after an uneventful pregnancy with no history of birth trauma. Her motor milestones were delayed: she was only able to sit by the age of 2 and was not able to walk until the age of 3 - 4 years. Her speech was also delayed, but initially appeared only mildly affected. At the age of 7, her speech began to deteriorate, becoming progressively difficult to understand, and this was followed by progressive deterioration in her walking, with increasing unsteadiness and falls. By the age of 11, she was no longer able to walk independently and required a frame. Spasms of the hands and feet began around the age of 15 and gradually worsened. On examination for this project at the age of 20, visual acuity was 5/12 in the right eye and 6/36 in the left. There was exotropia and hypotropia of the left eye along with latent nystagmus and dissociated vertical deviation. These eye signs were not present in her affected sister and were not felt by the neuro-ophthalmologists to be related to the dystonic condition but to represent a separate chance occurrence of a common and well recognised paediatric ophthalmic syndrome. Slit lamp examination revealed normal retinae. Speech was markedly dysarthric. There was generalised

dystonia affecting all four limbs, cranial region and trunk, with evidence of scoliosis. Choreoathetosis was also evident in the outstretched fingers. Power was generally preserved, with exception of some mild weakness of the first dorsal interossei, abductor pollicis brevis and finger extensors bilaterally. Sensation was normal. Coordination was difficult to assess in the presence of the limb dystonia, but there was probably mild appendicular ataxia. Reflexes were pathologically brisk throughout and the plantars were extensor bilaterally. The gait was dystonic and unsteady. Cognition appeared grossly normal, though formal neuropsychometry had not been undertaken.

She had been extensively investigated as a child by the neuro-paediatricians. Notably, metabolic profiles, including alpha-fetoprotein, lipoprotein A, white cell enzymes and very long chain fatty acids, were normal. Genetic testing for *TOR1A*, *THAP1* and *PANK2* was negative. CSF analysis, including oligoclonal bands, pterins and monoamine metabolites, was unremarkable. A DAT-scan revealed a normal pattern of tracer uptake. MRI of the brain, however, demonstrated cerebellar atrophy and bilateral pallidal hyperintensities on T2-weighted sequences felt to be consistent with limited bilateral striatal necrosis (figure 48).

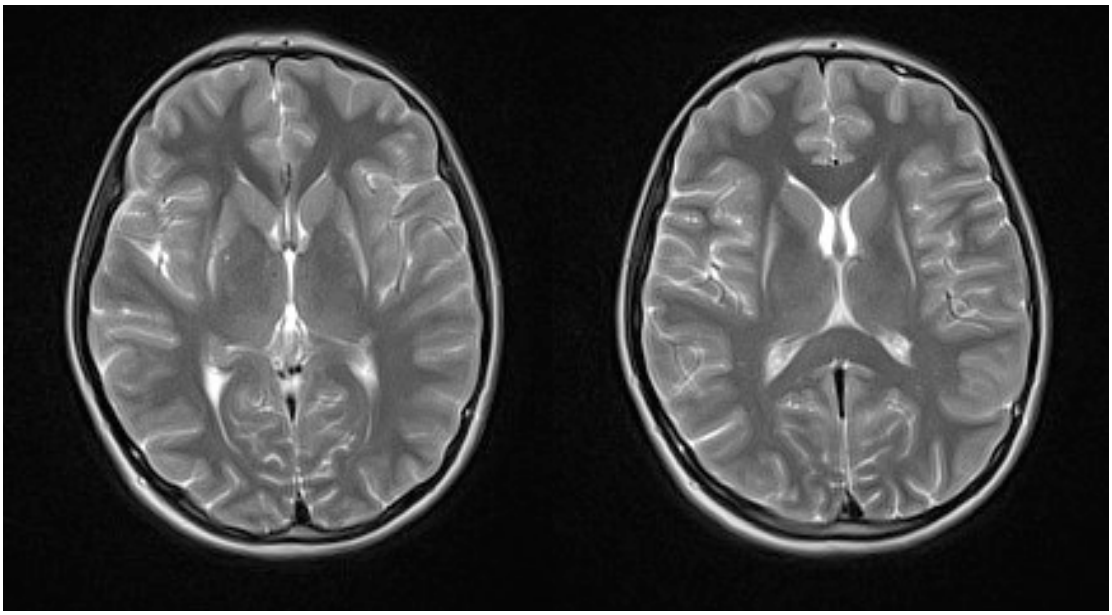

Figure 48 - MRI of the brain of individual II-1 demonstrates T2 hyperintensity in the region of the globus pallidus bilaterally, interpreted by the neuroradiologist to represent limited bilateral straital necrosis. Similar changes were evident on the MRI of her affected sister. Mild cerebellar atrophy (not shown) was also noted.

Individual II:3, who was two years younger, was also affected. The history of her illness had followed the same pattern as that of her sister, though the dystonia was slightly milder. Her examination at the age of 18 revealed the same neurological signs and deficits, with the exception that there was no exotropia, hypotropia, latent nystagmus, or amblyopia. In addition to the dystonia and choreoathetosis, occasional myoclonic jerks were also evident in the arms of this individual.

Individuals II:2, II:4 and both parents were neurologically normal.

DNA was available from all individuals shown on the kindred. Given the fact that both parents were neurologically normal in the context of the occurrence of a young-onset, relatively severe condition in two of their female children, the most probable mode of inheritance was autosomal recessive. Moreover, in view of the lack of any reported consanguinity, it was hypothesised that compound heterozygous mutations would be most likely to be responsible for disease in this family

10.2.2 Whole Exome Sequencing

DNA from both affected siblings and their mother was used to perform whole exome sequencing. Exome sequencing was performed using Illumina's TruSeq (62Mb) DNA Sample Prep Kit and Exome Enrichment Kit as per section 4.15. Reads were subsequently aligned and annotated using the standard approaches set out in sections 4.16. Details of the filtering of exome sequencing data are given in the results section below.

10.2.3 Genome-Wide Genotyping by DNA Microarray

Genome-wide genotyping data was obtained for all 4 siblings and for individual I:1 (the father) using the OmniExpress platform, which genotypes approximately 500,000 markers, as per section 4.12..

10.2.4 Autozygosity Mapping and Linkage Analysis

Despite the fact that the family did not report any consanguinity, autozygosity mapping, performed as per section 4.13, was employed for the sake of completeness. The extent

of homozygosity was quantified and tracts of homozygosity shared only by affected individuals were identified that could be further scrutinised for potentially causative homozygous variants.

Linkage analysis was performed as per section 4.14. Both a parametric analysis (using an autosomal recessive model with a penetrance of 100% and a pathological allele frequency of 0.0001) and a non-parametric analysis were performed to highlight areas that might habour a causal variant. Areas of interest were defined as those with a LOD score of greater than 0.

10.2.5 Sequence Analysis of Genes Known to be Associated with Bilateral Striatal Necrosis

In order to exclude the possibility that disease in this family was due to mutation in a gene already known to be associated with hereditary bilateral striatal necrosis, the sequence of these genes were scruntinised for possible pathogenic mutations. In the case of the autosomally-located genes, *SLC25A19*, *NUP62* and *NDUFV1*, the exome sequencing data was uploaded to the Integrative Genomics Viewer v2.2 (Broad Institute) for inspection. Each gene was first inspected to ensure all coding exons had been covered and subsequently the sequence of each exon in turn was inspected to look for variants detected. In the case, *ATP6* and *ND6*, which are encoded by mitochondrial DNA, Sanger sequencing was performed, as per section 4.6 – 4.10, using primers previously designed and used by our own diagnostics laboratory.

10.3 Results

10.3.1 Exome Sequencing

Exome sequencing produced excellent coverage in all three individuals. Using the CCDS hg19 definition of the exome, the average mean read depth across the entire exome was 55 with 93.5% covered by at greater than 2 reads, 81.3% covered at greater than 10 reads and 80.4% covered at greater than 20 reads. This translated to an average variant count of 22,250 variants per exome.

10.3.2 Autozygosity Mapping

Autozygosity mapping demonstrated that runs of homozygosity greater than 1Mb in length were both rare and small (maximum size < 3Mb), suggesting that recent consanguinity was not a feature within this family. Only three small runs of homozygosity that were shared by the two affected siblings and not present in the two unaffected siblings were identified (summarized in table 30).

Within these regions, *KCND2* (a potassium channel that contributes to repolarization in cardiac cells) is the only known gene annotated by Ensembl. Manual examination of the exome data for this gene showed that it had been well covered and no potentially-causal variants had been detected. In summary, this analysis suggested that the causal variants were more likely to be compound heterozygous changes, as previously hypothesized.

Table 30 - Details of the runs of homozygosity found to be shared between both affected siblings and not present in any unaffected sibling. Only one annotated gene was identified within these small regions, *KCND2*. Biological plausibility for this gene was low and inspection of the exome data showed it had been well covered without detection of any potential causal variants.

Chromosome	Start	Stop	Size (Kb)	Known Genes
3	95058076	96066172	1008	None
7	118229822	120273226	2043	*KCND2*
23	25798188	27120547	1322	None

10.3.3 Linkage Analysis

Linkage analysis was not expected to produce a truly significant LOD score, but was instead used to highlight areas with a greater *a priori* probability of harboring the causal variant. All peaks of positive linkage shared a common maximal LOD score of approximately 0.8 and were thus all treated as areas of potential interest. The position and size of these regions is shown in table 31.

Table 31 - Physical characteristics of linkage peaks detected in this family. Given the small size of the family, multiple peaks with a maximal LOD score of 0.8 were identified, which formed the primary areas of interest.

Chromosome	Start Position	Stop Position	Size (Mb)
1	37185336	91528185	54.34
2	23050710	31066592	8.02
	62430208	114273105	51.84
	139492011	170146913	30.65
3	72354511	111138405	38.78
	191860535	196452238	4.59
5	39087074	57411570	18.32
	95882162	101831523	5.95
6	153588258	167018721	13.43
8	4219026	13521103	9.30
9	107552924	108893850	1.34
10	570172	5593361	5.02
11	45051827	83282729	38.23
12	3255362	8068881	4.81
	46361672	66829726	20.47
13	27409049	30855014	3.45
14	24872917	33609861	8.74
15	37840246	85423958	47.58
18	2761338	11025325	8.26
20	129635	10968733	10.84
	60216060	61008537	0.79
22	36545780	46931838	10.39

Of note, there is no linkage peak over the chromosomal positions of *NUP62* (chr19, 50,410,082 - 50,432,988) or *SLC25A19* (chr17, 73,269,061 – 73,285,591), making it distinctly unlikely that these genes were responsible for disease in this family. *NDUFV1* is, however, covered by a linkage peak on chromosome 11.

10.3.4 Exclusion of Genes Known to Cause Bilateral Striatal Necrosis

Initially, the exome data was manually inspected to look for variants in *SLC25A19*, *NDUFV1* and *NUP62*. All exons and the intron/exon boundaries of all three genes were well covered in the data (> 20 reads minimum; often greater than 50 reads). No potentially causative variants were present (see table 32). Sanger sequencing of *ATP6* and *ND6* failed to reveal any variants in these genes.

Table 32 - Results of sequence analysis of known genes capable of causing bilateral striatal necrosis. *SLC25A19*, *NUP62* and *NDUFV1* were analysed via the exome data. *ATP6* and *ND6*, which are encoded by mitochondrial DNA, were sequence using Sanger methodology. Most variants detected were non-coding. The sole coding variant detected was a common SNP present in around 30% of the population, thus essentially ruling it out as a cause of disease in this family.

Gene	Position	Base change	Protein change	SNP number	MAF (dbSNP)	Zygosity
SLC25A19	73285461	C>A	None	rs2291033	0.11	hom
	73279624	G>C	None	rs7213318	0.09	het
	73269676	C>T	None	rs4789164	0.41	hom
NUP62	50412417	G>A	None	rs999583	0.77	hom
	50412217	C>G	S283T	rs79747934	0.34	hom
	50411742	A>G	None	rs892028	0.95	hom
NDUFV1	No variants detected by exome sequencing					
ATP6	No variants detected by Sanger sequencing					
ND6	No variants detected by Sanger sequencing					

10.3.5 Identification of Candidate Causal Variants by Appropriate Filtration

In order to isolate potentially pathogenic variants, we applied a systematic filtering procedure to the data, as depicted in figure 49. We began by selecting only those variants that were present in the exome data both affected sibling for analysis. Synonymous variants (unless in a splicing region) and variants recorded in dbSNP135 were initially removed. We then filtered out any variant present at a global minor allele frequency of ≥1% in a range of publically available databases of sequence variation

(1000 Genomes, Complete Genomic 69 Database and NHLBI Exome Sequencing Project database) as well as those found in 2 or more our own in-house exomes from individuals with unrelated diseases ($n \approx 200$).

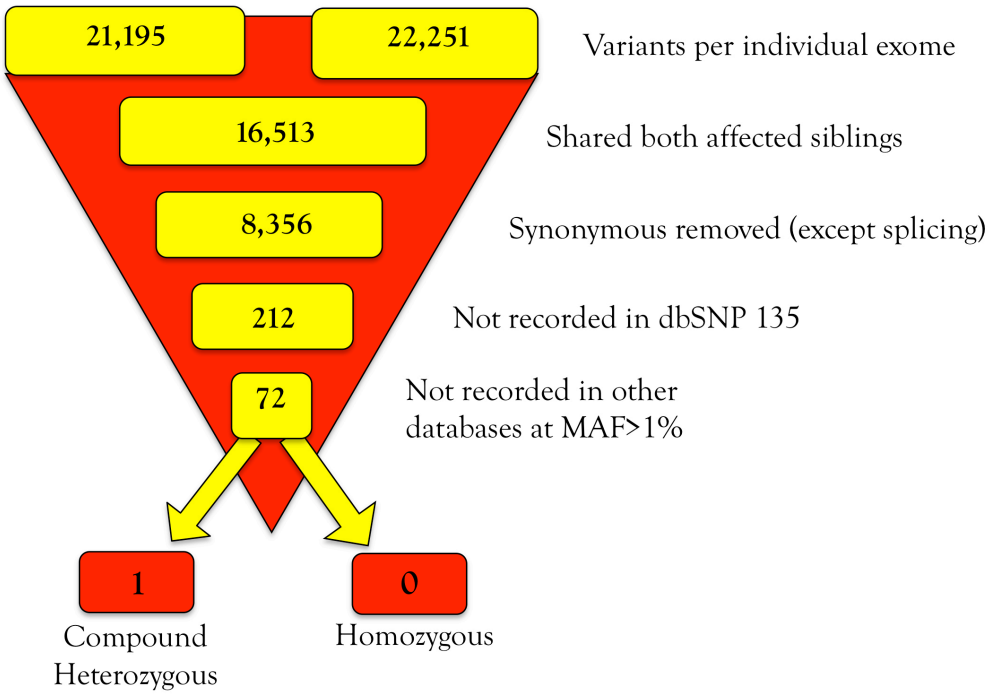

Figure 49 - Schematic representation of the filtration process employed to identify candidate causal variants for disease in this family, using all variants detected in the exomes of both affected siblings as starting position.

We identified only one gene that harbored potentially causative compound heterozygous mutations, *NUBPL*. This gene lies in the center of a linkage peak on chromosome 14. Using the longest transcript as a reference (ENST00000281081), the first variant is a T to C substitution at position 311 of the cDNA, resulting in the incorporation of proline in place of a leucine at position 104 of the protein product. It is predicted to be highly damaging by all four in silico prediction programs (SIFT=0; Provean=-6.5; PolyPhen2=1; MutationTaster=0.99). The affected base is highly conserved (PhyloP=3.09; GERP=5.4). The variant is reported in dbSNP137 with a minor allele frequency of 0.001 and has been detected once in the heterozygous state in the 11,779 chromosomes sequenced by the NHLBI Exome Sequencing Project.

Homozygosity for this variant would thus be expected to occur in approximately 1 in 4,000,000 births. It is one amino acid upstream of a previously reported pathogenic mutation in this gene (p.D105Y; see section 10.4.2) [410].

The second variant is A to T substitution at position 287 of the cDNA, resulting in the incorporation of a valine in the place of an aspartic acid at position 96 of the protein product. It is predicted to be tolerated by SIFT and Provean (though the Provean prediction is damaging for some transcripts), 'possibly damaging' by PolyPhen2 and damaging by MutationTaster (0.99). The affected base is highly conserved (PhyloP=2.228; GERP=5.41) and the variant is not found in any database of human variation.

10.3.6 Segregation Analysis in the Index Family

Inspection of the mother's exome data confirmed that she carried only the p.D96V mutation and that the two mutations were thus inherited on separate alleles in the affected siblings. Sanger sequencing was undertaken in the entire family for both variants, which demonstrated that they segregated with disease in the compound heterozygous state (figure 50). One of the unaffected siblings was heterozygous for the p.D96V mutation, while the other possessed two wildtype alleles.

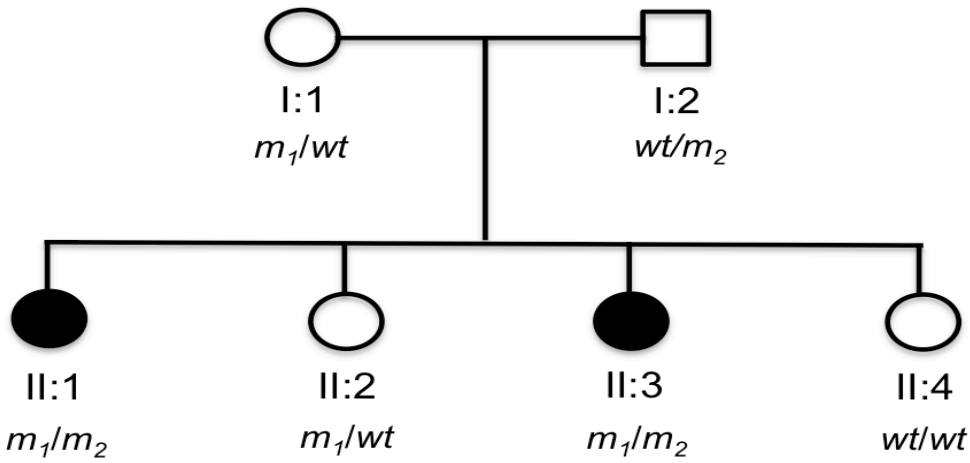

Figure 50 - Results of the segregation analysis for the two mutations in the gene NUBPL in the index family. Disease can be seen to segregate with the mutations in the compound heterozygous state. m_1 = p.D96V mutation; m_2 = p.L104P mutation; wt = wildtype allele.

10.3.7 Attempts to Generate Further Genetic or Functional Evidence That Mutations of NUBPL is the Cause of Disease in this Family

Given the identification that the mutations in *NUBPL* segregated with disease in this family, we queried our database of stored research samples for samples donated by other individuals with bilateral striatal necrosis in order to attempt to assemble a cohort of phenotypically similar individuals in which the *NUBPL* could be sequenced. Unfortunately, no sample could be found that included bilateral striatal necrosis in the clinical details. There are a number of reasons why this might be. Firstly, bilateral striatal necrosis is a very rare condition and only the largest medical institutions are likely to have more than one or two cases. Secondly, bilateral striatal necrosis is intimately linked to mitochondrial dysfunction and many individuals develop severe, widespread neurological disease. As a consequence, they often present in infancy and undergo investigation in a paediatric institution. Some do not survive to adulthood and those that do may not undergo any further genetic investigation meaning no DNA samples are available to us. Finally, some samples originate from individuals that were originally investigated at a time when MRI was used less routinely and neuroimaging may have consisted of a CT scan only, meaning that the presence of striatal necrosis may be missed.

As it would not be possible to generate further genetic evidence in the form of a second segregating family due to the lack of a cohort to screen, we gave thought to the possibility of functional studies to demonstrate pathogenicity of the detected mutations. Given that *NUBPL* encodes a protein that is required for the assembly of the respiratory chain NADH dehydrogenase (complex I), the simplest means of doing this within an acceptable time frame would have been to obtain patient fibroblasts and demonstrate that they showed evidence of complex I deficiency on bioassay. Unfortunately, neither of the two affected individuals wished to undergo any further invasive procedures, including a skin biopsy, and so it was not possible to pursue this means of establishing the pathogenicity of the mutations. Although it would be theoretically possible to engineer cells containing the two mutations by mutatgenesis and use these to test the effect of each mutation on complex I function, this is an extremely time-consuming process. Given that this would merely lead to the

identification of a new phenotype for a known gene rather than the discovery of a novel disease gene, this was not considered worthwhile in terms of the financial and research time commitment.

10.4 Discussion

We identified two potentially pathogenic variants in the gene *NUBPL* that segregated with disease in the compound heterozygous state in a family with autosomal recessive pallido-pyramidal syndrome and MRI evidence of limited bilateral striatal necrosis and cerebellar atrophy. Our attempts to generate the further evidence that would normally be required for publication of this finding failed for two reasons. Firstly, from a genetic point of view, no attempt could be made to find a second independent kindred with disease related to bilallelic mutations in this gene as there were no research samples available from phenotypically similar individuals for screening. Secondly, from a functional point of view, the hypothesis that fibroblasts from affected individuals harboring both mutations would show reduced complex I activity could not be tested as neither individual wished to undergo any further invasive procedures. Therefore, for the purpose of the following discussion, I have focused instead on a theoretical discussion of the *a priori* plausibility of mutations in *NUBPL* as a cause of dystonia with bilateral striatal necrosis in view of what is already known about the gene and the condition in question.

10.4.1 NUBPL is Connected to the Mitochondrial Respiratory Chain

NUBPL encodes a protein that is required for the assembly of the respiratory chain NADH dehydrogenase (complex I), an oligomeric enzymatic complex located in the inner mitochondrial membrane, consisting of 45 subunits and 8 iron-sulfur (Fe/S) clusters [411]. The protein encoded by *NUBPL* is an Fe/S protein that is thought to be involved in the transfer of Fe/S to the Fe/S-containing complex I subunits [411]. Its knockdown causes improper assembly of the peripheral arm of complex I, reduced complex I activity, and abnormal mitochondrial morphology [412]. It is notable that two other genes known to cause bilateral striatal necrosis, *ND6* and *NDUFV1*, are also components of complex I, whilst *ATP6* is a component of complex V, emphasizing the importance of mitochondrial bioenergetics in the aetiogenesis of this condition.

SLC25A19 encodes a mitochondrial thiamine pyrophosphate carrier and mutations in this gene are also likely to result in respiratory chain dysfunction.

10.4.2 NUBPL Has Previously Been Associated with Neurological Disease

Mutations in *NUBPL* have previously been associated with disease resulting from complex I deficiency. To date, 7 patients habouring biallelic mutations in NUBPL have been described in the literature [410]. These individuals all exhibited early onset of severe neurological disease with a clinical picture dominated by ataxia, dysarthria and variable spasticity. Only 2 of the 7 patients were ever able to walk without support and only then at 8 years and 11 years of age, respectively. None showed evidence of dystonia or any other evidence of extrapyramidal involvement. MRI scans undertaken in these individuals showed cerebellar atrophy and cerebellar subcortical white matter signal abnormalities in all, along with frequent brainstem signal abnormalities and cerebral deep white matter lesions (5 of 7 patients), predominantly affecting the frontal or frontoparietal regions. None showed evidence of basal ganglia abnormalities on MRI.

At first glance, therefore, although there is some overlap clinically (pyramidal dysfunction and dysarthria) and neuroradiologically (cerebellar atrophy) between these individuals and our family, there are significant differences including most notably the presence of generalised dystonia and basal ganglia signal abnormalities and the lack of any white matter abnormalities on MRI in our patients.

Let us presume for a moment that the biallelic mutations in *NUBPL* detected in this study are the cause of disease in our family. What explanations could be advanced to account for the difference in phenotype between our own family and that of individuals with disease due to mutations in this gene previously described in the literature? In fact, at least three possible explanations exist to account for these differences for these differences.

10.4.3 The Current Phenotype Associated with NUBPL Mutations is a Self-Fulfilling Prophecy Resulting from Systematic Ascertainment Bias

After the initial discovery of mutations of *NUBPL* as a cause of complex I deficiency, subsequent patients for mutational screening were selected using MRI pattern recognition from a database of more than 3,000 cases using 'unclassified leukoencephalopathy' as the search criteria [410]. Presumably this was done because it was realized that the initial case had a distinctive pattern of MRI abnormalities that closely resembled a pattern previously described by Wolf *et al.* as defining a subset of patients with isolated complex I deficiency [413], though this is not explicitly stated. According to subsequent publications that described new cases with *NUBPL* mutations, inclusion criteria for screening used were: 1) extensive cerebellar cortex signal abnormalities; 2) signal abnormalities in the corpus callosum; and 3) absence of signal abnormalities in the basal ganglia, thalami, and cerebral cortex. Thus, by definition of criterion 3, no individuals with MRI appearances similar to those of our own family would have been screened, meaning that even if basal ganglia signal abnormalities exist as part of the phenotypic spectrum of *NUBPL* mutations they could never have been identified by these studies.

10.4.4 Functional Data Suggest the Currently Recognised Phenotype May Be the Most Severe End of the Spectrum

Secondly, all previously reported individuals carry the same putatively causative genetic mutation on one allele, that is a branch-point splicing mutation intron 9 (c.815-27T>C). Interestingly, this is always found in combination with a missense mutation in exon 2 (c.166G>A; p.G56R) on the same allele [410]. Previous functional work involving the transgene expression of mutant *NUBPL* carrying the p.G56R alone showed that this missense mutation does not impact substantially on *NUBPL* mRNA or protein stability, protein import in mitochondria, or protein function [414]. Therefore, the c.815-27T>C branch-site mutation is thought most likely to be the cause of the complex I deficiency, although an additive effect of the missense mutation has not be excluded. RT-PCR further demonstrated that the branch-site mutation resulted in a profound decrease in the steady-state level of *NUBPL* mRNA, quantified as 15% of control levels. Two additional aberrant transcripts were observed, of which one is degraded by nonsense-mediated decay. The other aberrant transcript lacked exon 10 and exon 9 was directly fused to exon 11, causing a frameshift after glycine 272, with an additional 31 codons

until a stop codon is reached. As a consequence, aspartate 273 is substituted by a glutamine followed by 30 random amino acids resulting in an altered C-terminus (p.D273Q*fs*X31). This mutation leads to a decrease in *NUBPL* protein level [410, 414]. Finally, the equivalent mutant protein in the yeast *Yarrowia lipolytica* (Ind1p.N271QfsX31) has been shown to be completely non-functional and does not support the assembly of complex 1 [415]. In about half of the 7 reported cases, this pathogenic branch site mutation was inherited with a presumed null allele. Thus, given the above data, at least half of the currently described cases would be expected to have almost no functional *NUBPL* activity whatsoever. It is quite possible, therefore, that the individuals so far reported represent the most severe end of a spectrum of disease that mutations in *NUBPL* may be capable of causing. Other mutations in this gene, such as those that we identified in this study, may be associated with a wider spectrum of disease than has so far been recognised, including much milder phenotypes where extensive leucoencephalopathy is not a feature.

10.4.5 Phenotypic Heterogeneity is Common in Disorders Involving the Mitochondria

Thirdly, diseases in which mitochrondrial dysfunction is key to pathogenesis tend to show a much wider spectrum of phenotypes than diseases due to dysfunction in other cellular organelles or pathways. In the case of disease resulting from the mutation of genes encoded by mitochrondrial DNA, this is readily explained by mitochondrial heteroplasmy within the oocyte leading to a variable load of mutatated mitochrondrial DNA in the cells of the progeny. However, the variability in the severity of disease resulting from mitochondrial dysfunction extends beyond this and is also evident when the disease gene is an autosomally-encoded mitochondrial gene. For instance, the gene *SLC25A19*, located on chromosome 17, has been shown to cause bilateral striatal necrosis [409]. The affected individuals, who all carried a homozygous p.G125S mutation, exhibited onset of lower motor neuron weakness in the first decade of life, leading to atrophy and contractures. On nerve conduction studies an axonal neuropathy was evident and MRI of the brain demonstrated bilateral T2 hyperintensities in the striatum suggestive of necrosis. Cognitive abilities were unaffected and no pyramidal, extra-pyramidal or cerebellar signs were detected on examination. However, mutations

in *SLC25A19* had already been described as causing disease in patients of Amish origin, who all harbored a different homozygous mutation of this gene (p.G177A) [416]. These individuals universally exhibited a much more severe phenotype with microcephaly, onset of weakness by the 20th gestational week, failure to obtain any developmental milestones and, finally, death in infancy. Likewise, mutations in *NDUFV1*, another autosomally-encoded gene associated with bilateral striatal necrosis, are also associated with much more severe phenotypes, including diffuse leucoencepahopathy [417] or Leigh syndrome with or without myoclonic epilepsy [418, 419]. Thus, knowledge of other mitochrondrial disorders suggests that a wide phenotypic spectrum of disease is common. It all probability the severity of the phenotype is determined by the functional impact of particular mutations on the mitochrondrial respiratory function, which, given their central role in cell survival, can translate into disease phenotypes that range from mild to lethal. Moreover, extensive leucoencephalopathy, as present in all currently described cases of *NUBPL* mutation carriers, is recognised as the more severe end of the spectrum of disease resulting from other genes that have also been shown to cause limited bilateral striatal necrosis.

10.4.6 Summary

Although it was not possible in this case to prove by the standards usually required that the biallelic mutations in *NUBPL* were the cause of disease in this family, the lack of any other candidate varaints on exome sequencing and the inherent biological plausibility of the implicated gene within the framework of what is currently known about other genetic causes of bilateral striatal necrosis is persuasive. Moreover, although *NUBPL* mutations have already been associated with neurological disease that is somewhat different and certainly more severe than that seen in our family, I have set out several good reasons why this cannot be used to rule out *NUBPL* as a cause of disease in the kindred presented herein.

CHAPTER 11:

Exome Sequencing in Severe, Infantile-Onset, Autosomal Recessive Choreodystonia

11. Exome Sequencing in Severe, Infantile-Onset, Autosomal Recessive Choreodystonia.

11.1 Introduction

Complex neurological phenotypes, including variable degrees of dystonia, have been associated with recessive mutations in a number of genes (see table 9 for examples). However, many complex phenotypes resist diagnosis and a proportion of these will result from mutations in Mendelian disease genes that are not yet known.

We identified a consanguineous Pakistani kindred in which two of the four siblings were affected by an undiagnosed, infantile-onset form of severe choreodystonia. The phenotype was not recognised as typical of any known genetic disease by the movement disorders physicians at the National Hospital for Neurology and Neurosurgery and previous mutational screening of genes associated with both isolated and complex degenerative dystonic syndromes had all been negative (see section 11.2.1 below).

The family were originally based in Scotland, but had recently moved to London and had come under the care of the movement disorders physicians at the National Hospital for Neurology and Neurosurgery. They agreed to participate in our genetic research to attempt to elucidate the cause of the disease in their family and signed the necessary consent forms. Unfortunately, shortly after an initial brief meeting with the family to obtain samples of blood for genetic analysis and examine the core family, one of the affected siblings died quite suddenly and unexpectedly.

At this point, although the family reaffirmed their consent for us to continue with the genetic research based on the samples we had already obtained, they asked that they not be contacted again until they had had time to come to terms with their loss. We provided them with contact details so that they might re-establish contact when they felt ready. However, at the time of writing, we had not heard from them again and, in line with their wishes, we have not made any further attempt at contact ourselves. We subsequently learned that they had returned to Scotland and are thus no longer under follow up at the National Hospital.

An unfortunate consequence of this unexpected loss of contact was that we were not able to establish the full genetic pedigree beyond that of the core family, which had been provided at our initial meeting, and no record of this existed in their brief medical notes. However, as we already knew that the parents were first cousins (and had been informed the grandparents on one side were also first cousins), we felt confident that the likely mode of inheritance was autosomal recessive and that the causal variant would thus be homozygous. Moreover, the family had stated that they were not aware of any other affected individual within their extended family and so lack of knowledge of the full genetic pedigree was not expected to impact on study design or limit results.

We therefore set about elucidating the cause of disease in this family using a combination of whole exome sequencing and autozygosity mapping. Unlike as had been possible for the kindred presented in chapter 9, no assumptions could be made regarding the likely cellular function of the causal gene.

11.2 Subjects, Materials and Methods

11.2.1 Clinical Details of the Index Family

Participants were drawn from a small Pakistani kindred. The pedigree of the core nuclear family is shown in figure 51. DNA was obtained from all individuals shown in the tree except individual II:2 who did not wish to participate in the study as he was still living in Scotland. He was offered the opportunity to provide a DNA sample by posted 'spit kit' to minimise intrusiveness, but declined.

There were at least two loops of consanguinity within the family, in that the affected children's mother and father were first cousins and one set grandparents were also first cousins. However, for the reasons set out in the section 11.1 above, we were not able to establish the structure of the genetic pedigree beyond the core nuclear family. We were, however, told that the family were unaware of any other individual within the extended family with neurological disease.

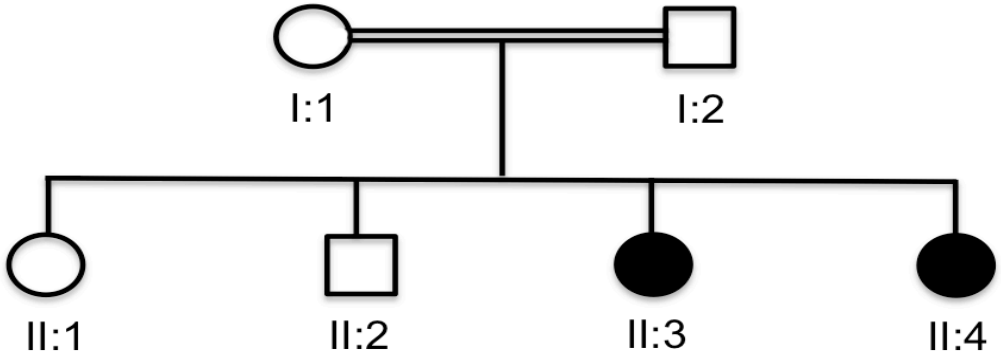

Figure 51 – The genetic pedigree of the index family. At least two consanguineous matings had occurred in this family: the first, shown here, was between individuals I:1 and II:2, who were first cousins; the second was between the grandparents on one side of the family, who were also first cousins. Unfortunately, for the reasons set out in section 11.1, we were unable to extend the genetic pedigree beyond what is shown here. The family did, however, inform us that they were not aware of any other individual in the extended family with neurological disease of any form.

Individuals II:3 and II:4 had both developed neurological signs in their infancy. They showed delayed motor milestones, but did temporarily achieve ambulation and remained able to speak at the time of their inclusion in this study. Their neurological illness was dominated by extrapyramidal features, in the form of dystonia with choreic elements, but the affected individuals also exhibited signs of cerebellar dysfunction, in the form of appendicular ataxia and dysarthria. At the time of inclusion in this study, individuals II-3 and II-4 were aged ?? and ??, respectively, and both required full time care and were confined to a wheelchair. Despite this, there were no pyramidal features and cognition was intact on bedside testing, though formal neuropsychometry had not been undertaken. MRI in one of the individuals showed cerebellar atrophy.

Both parents (individuals I-1 and I-2) and both older siblings (II-1 and II-2) were well, with no evidence of neurological disease (only the parents and individual II-1 could be examined as individual II-2 did not wish to participate).

DNA from the affected individuals had previous been screened for mutations in *TOR1A*, *PANK2*, *SCA1*, *SCA2*, *SCA3*, *SCA6* and *SCA7* and had tested negative.

11.2.2 Whole Exome Sequencing and Exome Coverage Statistics

DNA from both affected siblings was used to perform whole exome sequencing using Illumina's TruSeq (62Mb) DNA sample preparation and exome enrichment kits as per section 4.15. Reads were subsequently aligned and annotated using the standard approaches set out in sections 4.16. Coverage statistics for the exome as a whole were obtained using Picard.

11.2.3 Genotyping, Autozygosity Mapping and Coverage of Homozygous Regions

Genome-wide genotyping data was obtained for all three affected siblings using the OmniExpress platform, which utilises approximately 500,000 markers as per section 4.12 Autozygosity mapping was performed as per section 4.13.

Coverage across regions of shared homozygosity was calculated using BedTools against the CCDS definition of the exome and expressed as the percentage of bases target covered by at least one read. Briefly, using Ensembl's Biomart function, a list of every exon of every CCDS gene within each homozygous region was obtained, along with its start position, end position and strand. This information was used to construct a simple BED file of the desired targets within the region. Coverage of these target regions was then checked against the BAM file for the exome using BedTools 'coveragebed' function.

11.2.4 Filtration of Variants Detected by Exome Sequencing

In view of the apparently recessive inheritance pattern and history of consanguinity, we initially selected all homozygous variants for consideration. Subsequently, synonymous variants not predicted to affect a canonical splice site (i.e. those not within 10 bases, in either direction, of the intron/exon boundary) were discarded. Given the rarity of the disorder present in this family, we further hypothesised that the causal variant was unlikely to be found in any database of normal sequence variation. However, in order

to minimize the possibility of incorrectly assigning causality, we filtered out those variants that were recorded in the 1000 Genomes, NHLBI or Complete Genomics 69 datasets at a minor allele frequency of greater than 0.5%. Given that a variant found at even this low frequency would be expected to be occur naturally in the homozygous state in around 1 in every 160,000 births, it seemed distinctly unlikely that we would risk filtering out the causal variant with this cut-off. Subsequently, variants that were located in a region of shared homozygosity were selected as potentially causal. No filtration was performed on the basis of *in silico* predictions of pathogenicity or conservation scores.

11.2.5 Generation of Nucleotide Multispecies Protein Alignments

In order to assess conservation of amino acid sequence, Uniprot, FlyBase and WormBase were interrogated for orthologous protein sequences. Actual alignment of the sequences obtained in this manner was performed using the freely available web-based application ClustalΩ.

11.2.6 Generation of Regional Gene Expression Data

Information on organism wide and brain region specific gene expression was collated as per section 4.19.

11.3 Results

11.3.1 Exome Sequencing

Exome sequencing produced excellent coverage in all three individuals. Using the TruSeq definition of the exome, the average mean read depth across the entire exome was 42 with 93.5% of the target covered by at greater than 2 reads, 88.3% covered at greater than 10 reads and 80.4% covered at greater than 20 reads. This translated to an average variant count of 22,250 variants per exome.

11.3.2 Autozygosity Mapping

Autozygosity mapping demonstrated that runs of homozygosity greater than 1Mb were relative common in three siblings, with the largest single run of autosomal homozygosity stretching for 60Mb, confirming the report of extensive consanguinity

within the family. The runs of homozygosity shared between all both affected siblings but not present in the unaffected sibling are summarized in the table 33.

Initially, coverage of the homozygous regions was checked as described in section 11.2.3. Subsequently, the number of variants of all types detected in each homozygous region was determined along with the number of variants that might be potentially causal (i.e. those remaining after filtration as set out in the section 11.2.4 above). This information is also summarized in the table 33.

Table 33 - Summary of homozygous regions shared between both affected individuals but not present in their unaffected sibling, showing chromosomal location, size, percentage coverage, total variants contained therein and potentially causal variants contained therein.

Chr	Start	End	Length (Mb)	CCDS Genes	% Coverage	Variants Detected	Potentially Causal
3	47598663	49590770	1.99	50	88.9%	28	0
3	88708334	89097831	0.39	0	-	0	0
5	3205018	7859434	4.65	12	88.9%	7	0
5	72601498	73950760	1.35	6	82.3%	1	0
6	95090000	97751948	2.66	8	97.8%	3	0
6	163721073	170382923	6.66	28	87.3%	37	0
7	67165048	72283565	5.12	3	90.5%	0	0
7	94900975	103672960	8.77	127	88.1%	59	0
7	118356108	120278563	1.92	1	100%	0	0
7	137435711	139924127	2.49	19	86.6%	22	0
8	85152853	86181870	1.03	5	89.9%	0	0
11	3028140	12286355	9.26	151	90.3%	239	4
13	35135179	39793979	4.66	22	93.2%	17	0
13	88799824	89875736	1.08	0	-	0	0
14	91828769	92095855	0.27	3	96.3%	0	0

As detailed above, all homozygous regions demonstrated coverage of coding bases within the range expected for exome sequencing (>80%). In particular, the largest

region of homozygosity on chromosome 11, which haboured the only potentially causal variants, was 90% covered.

11.3.3 Overview of Potentially Causal Variants

After appropriate filtration as set out in section 11.2.4, four homozygous variants were identified that met the criteria for being potentially causative, all located on chromosome 11. These variants are summarized in table 34 and discussed individually in the following sections.

Table 34 – Details of potentially causal variants remaining after filtration of exome data, including chromosomal location, sequence change, amino acid substitution, zygosity, novelty, minor allele frequency (if previously observed) and number of reads covering the variant.

Chr	Position	Gene	Change	Zygosity	Novel?	MAF	Read depth
11	3381741	*ZNF195*	c.497C>T p.A169V	Hom	Yes	-	60
11	5510682	*OR52D1*	C746G>A p.G249D	Hom	No	0.001	70
11	6479001	*TRIM3*	c.440G>C p.G147A	Hom	Yes	-	10
11	6645261	*DCHS1*	c.7646G>A p.E2549R	Hom	Yes	-	22

ZNF195:

ZNF195 encodes a protein belonging to the Krueppel C2H2-type zinc-finger protein family. These family members are transcription factors that are implicated in a variety of cellular processes. No further information is available regarding its specific functions.

The variant identified by exome sequencing in this family is a C to T substitution at position 497 of the cDNA (ENST00000399602) resulting in the incorporation of a valine in place of an alanine at position 166 in the protein product. The affected base

is poorly conserved (PhyloP = -0.461; PhastCons = 0.406). In many species, this portion of the protein does not have a homologue, the two closest exceptions being the Chimpanzee and Rhesus Macaque, both of which also have an alanine at this position. The variant is not located in a key domain of the protein according to Uniprot, but resides instead in spacer region (amino acids 76 to 243) between an initial KRAB domain and the first of 10 zinc fingers. It is predicted to be tolerated by SIFT (0.245), just neutral by Provean (-2.201), probably damaging by PolyPhen2 (0.935), and a polymorphism by MutationTaster (0.999).

Publically available organism wide expression data for *ZNF195* shows that it is widely expressed throughout the body, including the brain (figure 52, left side). Within the brain it demonstrates relatively good expression in all areas sampled (figure 52, right side).

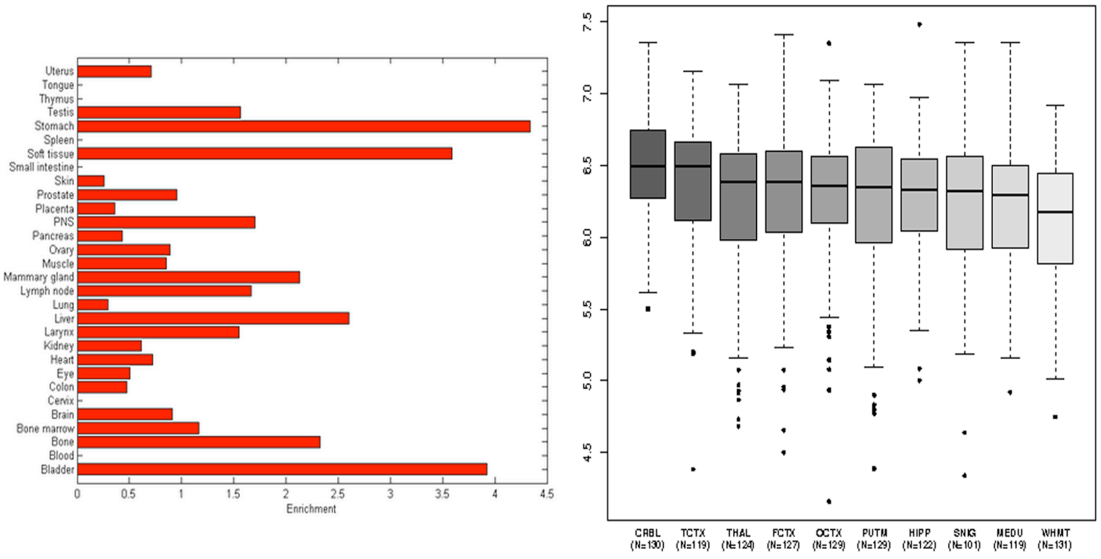

Figure 52 - Expression data for the gene *ZNF195* in man. The panel on the left shows the organism wide expression data based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 11.2.5, across the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior

olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

OR52D1

OR52D1 is a class I (that is, phylogenetically derived from aquatic animals) olfactory receptor (OR). Class I ORs were thought to be non-functional genetic relics, but recent work involving human *OR52D1* has demonstrated that it is, in fact, capable of generating currents in response to odorant stimuli, exhibiting a narrow repertoire related to that of its orthologous murine OR [420]. ORs are embedded in the plasma membrane of the olfactory neurons located in the olfactory epithelium and so do not appear a good candidate for mediating widespread neuronal dysfunction. At the same time, however, there is an emerging body of evidence to suggest ORs are expressed in other regions of the nervous system beyond the peripheral sense organs, including in the neurons of the cerebral cortex and other regions of the adult human brain [421]. It has been tentatively suggested that they many act as 'novel' chemoreceptors, guiding neuronal physiological processes in response to ligands transported by the blood, released by neighbouring cells or even by the same cell [422]. Certain "ectopic" ORs have been implicated in cell assembly functions in the embryonic period, muscle regeneration and regulation of cell adhesion and migration, and sperm chemotaxis [423-428]. In the CNS, axonal guiding through ORs has also been shown in the olfactory system [429-431]. Thus, although a seemingly unlikely cause of neurological disease, *OR52D1* cannot be dismissed on the basis of its being an olfactory receptor alone.

The variant detected by exome sequencing in this family is a G to A substitution at position 746 of the cDNA (ENST00000322641), which leads to the incorporation of an aspartic acid in place of a glycine at position 249 of the final protein product. The affected base is particularly well conserved (PhyloP=1.21; PhastCons=0.13). The variant is located in the sixth of seven predicted transmembrane helical domains, which begins at amino acid 241 and ends at amino acid 261. It is recorded in dbSNP (rs147388113) and present in 12 out of 12,996 chromosomes sequenced by the NHLBI Exome Sequencing Project, making it an unlikely cause of the disease in this family. It is

predicted to be damaging by SIFT (0.004), neutral by Provean (-1.423), possibly damaging by PolyPhen2 (0.751), and a polymorphism (0.999) by MutationTaster.

Of note, there is a second independent missense SNP (rs710919) just six bases downstream of the variant that we detected that leads to an isoleucine-to-threonine substitution at position 251 (i.e. within the same transmembrane domain). Despite being predicted to be more highly damaging by *in silico* methodology than the variant we identified by exome sequencing, this second SNP is common in the general population. It was present on 3,192 of 12,996 chromosomes sequenced by the NHBLI Exome Sequencing Project and was observed in the homozygous state in about 7% of individuals.

No EST expression data is available for this gene, but our own in house expression data shows low-level expression across the brain (not shown).

TRIM3:

TRIM3 encodes a protein that is a member of the tripartite motif (TRIM) family, also called the 'RING-B-box-coiled-coil' (RBCC) subgroup of RING finger proteins. The TRIM motif includes three zinc-binding domains, a RING, a B-box type 1 and a B-box type 2, and a coiled-coil region.

It is intimately involved with neuronal physiology, as is reflected in its original name, brain expressed ring finger protein (BERP) [432]. It is similar to a rat protein that is a specific partner for the tail domain of myosin V, a class of myosins, which are involved in the targeted transport of organelles. The rat protein can also interact with alpha-actinin-4. Evidence suggests that the human protein may play a role in myosin V-mediated cargo transport [433]. *TRIM3* deficient mice, which are viable, healthy and fertile with no gross of histological abnormalities, exhibit reduced synaptic $GABA_A$ receptor function, which was thought most likely to be due to disrupted kinesin 21B-dependent trafficking of the protein [434, 435]. Thus, a mechanism exists by which mutation of *TRIM3* could be envisioned to lead to aberrant neuronal excitability, which has been proposed a mechanism underlying dystonia. In addition to the above,

the has also been linked to nerve growth factor-induced differentiation and neurite outgrowth as well as synaptic remodelling, in the latter case as the ubiquitin ligase responsible for regulated degradation of key postsynaptic density proteins [436, 437]. A recent shotgun proteomics analysis has indicated that *TRIM3* is up-regulated in the brains of schizophrenic patients [438].

The expression data for *TRIM3* is also interesting with respect to the disease phenotype seen with in this family. As shown in figure 53, TRIM3 is highly expressed in the brain and shows particularly high levels of expression in the cerebellum. This is interesting for two reasons. Firstly, the only other significant neurological signs, besides dystonia, in the members of this family affected by disease were dysarthria and mild appendicular ataxia, suggestive of cerebellar dysfunction and their MRI showed cerebellar atrophy. Secondly, evidence from research on one of the best animal models of dystonia, the *dt* rat, suggests that the dystonic phenotype in this animal arises due to cerebellar dysfunction and, more precisely, as a result of abnormalities of GABAergic signalling in this structure. The *dt* rat develops progressively more severe generalized dystonia from around postnatal day 14, caused by a deficiency of the neuronally-restricted protein caytaxin [439]. Like TRIM3, caytaxin also binds to kinesin light chains to mediate intracellular transport of specific cargos [440] and, like TRIM3, it is predominantly expressed within the cerebellum [441]. Whilst no differences have been demonstrated in cell morphology in the basal ganglia of these rats, Purkinje cells in the vermian and paravermian tissues at post natal day 20 at around 5 – 11% smaller in *dt* rats than in their non-mutant littermates [439]. Moreover, levels of the second messenger protein, cGMP, are significantly reduced in *dt* rats and show smaller rises in response to harmaline challenge [439]. This suggests a deficit in neurotransmission in the Purkinje fibres of this rat, which is probably postsynaptic. Subsequent work has shown that the underlying deficit is increased activity of Purkinje cell GABAergic synapses within the *dt* rat cerebellar nuclei [442-444]. Thus, both TRIM3 and caytaxin have been linked to GABAergic activity in the cerebellum and caytaxin deficiency leads to dystonia.

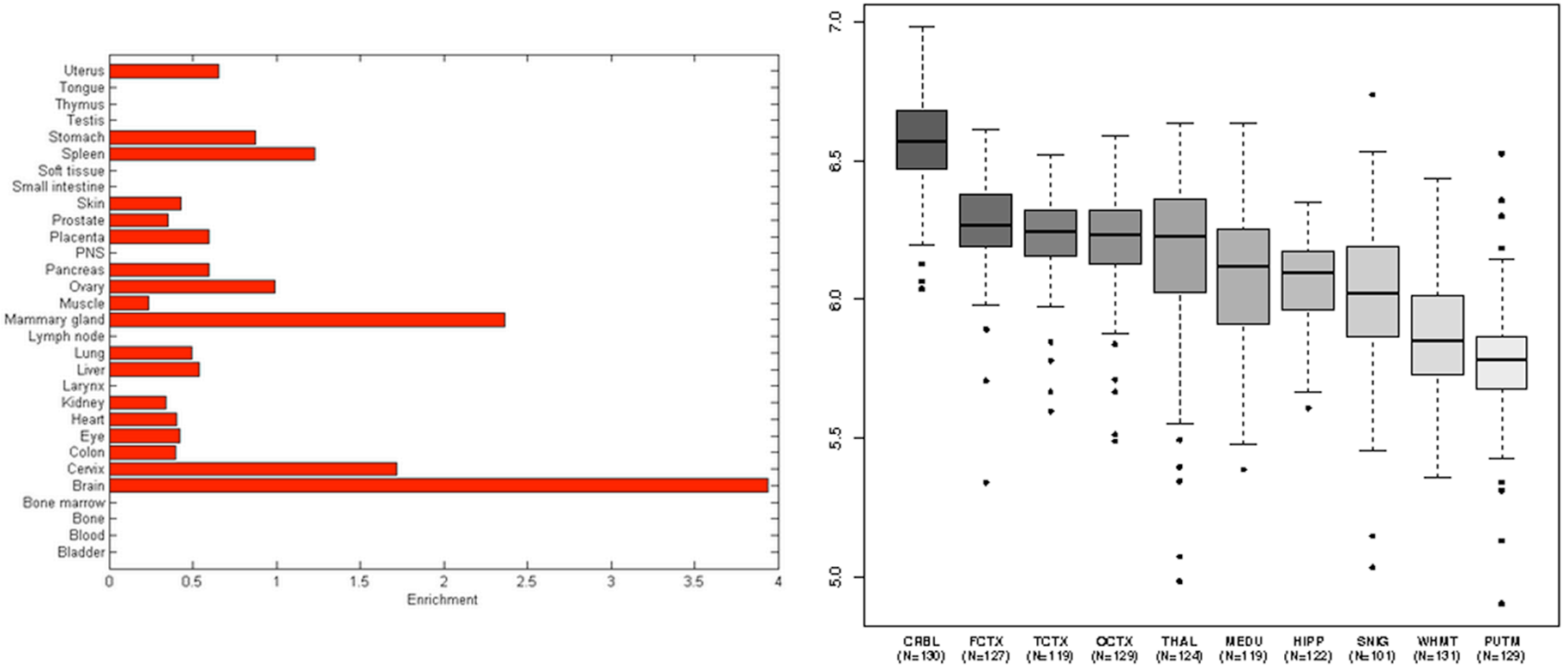

Figure 53 - Expression data for the gene *TRIM3* in man. The panel on the left shows the organism wide expression data based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 11.2.5, across the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

The variant detected by exome sequencing in this family is a G to C substitution at position 440 of the cDNA (ENST00000345851), which leads to the incorporation of an alanine in place of a glycine at position 147 of the final protein product. The affected base has a relatively neutral conservation score (PhyloP=0.398; PhastCons=1). The variant is, however, located in the key type 2 B-box domain of the protein, which coordinatively binds zinc ions through regularly spaced, ultra-highly conserved histidine and cysteine residues and is involved in ubiquitination. The amino acid affected by this mutation sits directly next door to one of these key histidine residues (amino acid 146). The B-box domain spans from amino acid 110 to 151.

Despite the critical location of the mutation, it is predicted to be tolerated by SIFT (0.903), neutral by Provean (0.307), benign by PolyPhen2 (0), and a polymorphism (0.644) by MutationTaster. At least part of the benign nature of these predictions undoubtedly owes to the fact that alanine, like glycine, is a small neutral amino acid and thus substitutions of these two amino acids are generally considered less likely to be deleterious to the protein. That said, glycine is a very unique amino acid in that it contains a hydrogen at its side chain (rather than a carbon as in the case of all other amino acids). This means that there is much more conformational flexibility in glycine, which allows it to reside in parts of proteins structures that are forbidden to all other amino acids (such as tight turns in structures). It can also play a distinct functional role by using it sidechain-free backbone to bind phosphates. In the final analysis, however, it is difficult to sustain the argument that TRIM3 function is reliant on the special properties of this particular glycine since a glycine is not seen at this position in a number of other species, including the chicken (G. Gallus), the tasmanian devil (S. Harrisii), the zebra fish (D. Reiro), the pufferfish (T. Rubripes), the nematode worm (C. Elegans), whereas the critical neighbouring histidine is conserved in all species (figure 54)

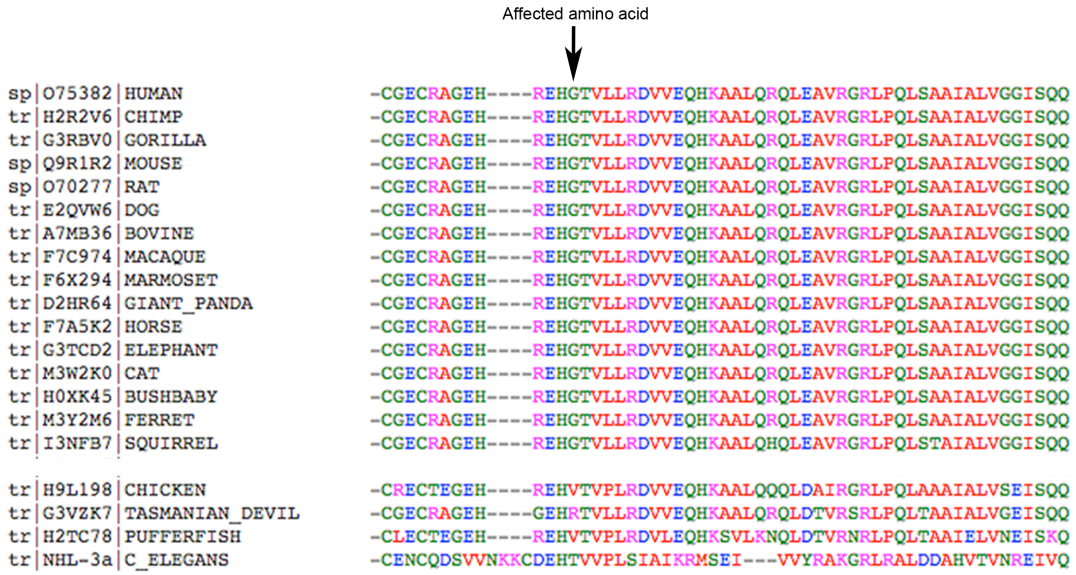

Figure 54 - Multiple species alignment shows only partial conservation of the affected amino acid in TRIM3. Conversely, the neighbouring histidine residue, which is known to be critical for TRIM3's function as a ubiquinase, shows absolute conservation in all species. Colours indicate physiochemical properties of amino acids (red = small/hydrophobic; blue = acidic; magenta = basic; green = hydroxyl/sulfhydrl/amine/glycine).

DCHS1:

DCHS1, standing for 'dachsous 1 (Drosophila)', is a member of the cadherin superfamily. This family of genes encode calcium-dependent cell-cell adhesion molecules. The particular protein encoded by *DCHS1* has a signal peptide, 27 cadherin repeat domains and a unique cytoplasmic region. It is expressed in fibroblasts and where it may play a part in wound healing. More relevant for our purposes, *DCHS1* also forms part of a pathway that regulates planar cell polarity in both Drosophila and mice and is involved in genesis of the mouse embryonic cerebral cortex [445, 446]. Transgenic mice carrying a copy of *Dchs1* with a targeted excision of the second exon (which prevents its expression by removing the initiator methionine, the signal peptide and first 6 cadherin domains) are viable at birth but fail to grow. They show multiple morphological abnormalities, including a widened neural tube, small cystic kidneys, a reduced intestinal length, smaller lungs, defects in atrial septation, skeletal abnormalities and curly tails [447]. This suggests that *DCHS1* is important in the

morphogenesis of multiple organs, in mouse at least, and pathogenic mutations in this gene might be expected to lead to a phenotype with extra-neurological manifestations.

The variant detected by exome sequencing in this family is a G to A substitution at position 7646 of the cDNA (ENST00000299441), which leads to the incorporation of an arginine in place of a glutamine at position 2549 of the final protein product. The affected base has a relatively neutral conservation score (PhyloP=0.133; PhastCons=0.911). The variant is located in the 24th cadherin domain, which begins at amino acid 2483 and ends at amino acid 2602. It is predicted to be tolerated by SIFT (1), neutral by Provean (0.159), benign by PolyPhen2 (0), and a polymorphism (0.999) by MutationTaster.

Publically available systemic expression data for *DCHS1* shows that it is widely expressed throughout the body, including at relatively high levels in the brain (figure 55, left side). Within the brain, it shows a particular emphasis on some areas associated with movement disorders, such as the striatum, thalamus and substantia nigra (figure 55, right side)

11.3.4 Selection of a Candidate Variant for Further Sequencing

None of the variants identified in this family are clearly causal. Those variants with a proven role in the brain (*DCHS1* and *TRIM3*) are also, unfortunately, the same variants that are predicted to be benign by all *in silico* programs. However, after careful consideration, it was decided that *a priori* biological plausibility of neuronal dysfunction as a result of disruption of TRIM3 function was sufficiently high to overlook the *in silico* predictions and proceed to a screening of the affected exon in a cohort of individuals affected with generalised dystonia in the hope of finding a second family with mutations in this gene.

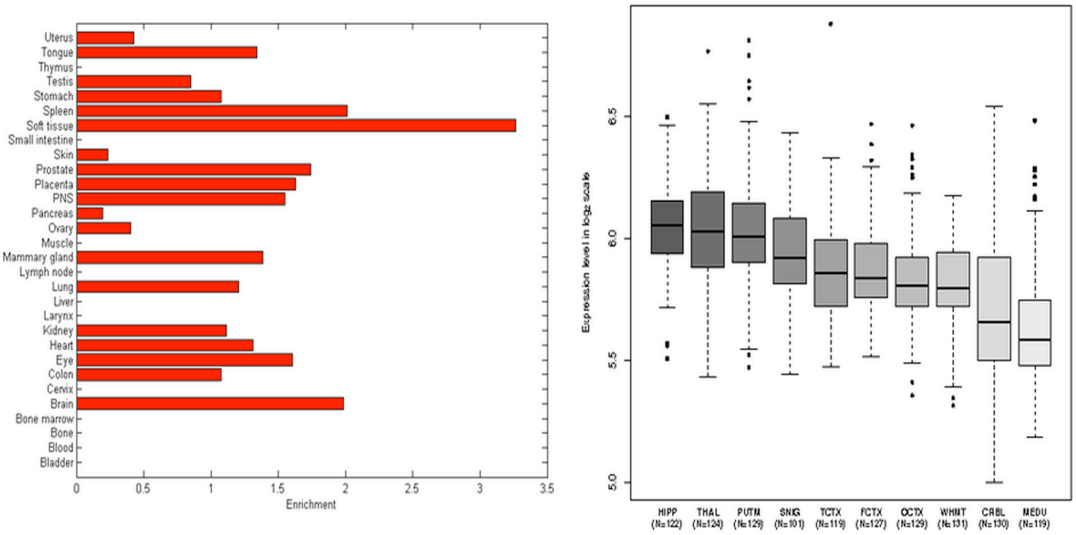

Figure 55 - Expression data for the gene *DCHS1* in man. The panel on the left shows the organism wide expression data based on publically available EST datasets. The panel on the right shows regional differences in expression across the brain, derived as described in section 9.2.6, across the 10 CNS regions analysed: putamen (PUTM, n=129), hippocampus (HIPP, n=122), temporal cortex (TCTX, n=119), frontal cortex (FCTX, n=127), occipital cortex (OCTX, n=129), thalamus (THAL, n=124), cerebellar cortex (CRBL, n=130), substantia nigra (SNIG, n=101), intralobular white matter (WHMT, n=131) and medulla (specifically inferior olivary nucleus, MEDU, n=109). Whiskers extend from the box to 1.5 times the inter-quartile range.

11.3.5 Screening of Exon 4 of TRIM3

Screening of exon 4 of *TRIM3* was undertaken using Sanger Sequencing methodology. 384 DNA samples belonging to individuals diagnosed with dystonia were selected from an in-house library of research samples. As was discussed in chapter 7, no samples were available where autosomal recessive inheritance of dystonia was definite. Therefore, samples were selected instead on the basis that the clinical details suggesting either no family history or a family history consisting only of affected siblings. Precise age at onset data is not available for most samples, but an attempt was made to ensure young-onset cases were screened by including those samples that showed a date of receipt within 30 years of the donner's birthdate. All samples had previously been screened for mutations in *TOR1A* and were negative.

No previously unreported, potentially causal mutations were identified in our cohort. The only novel change detected was a heterozygous synonymous SNV (c.384G>A) which is located 20 bases from the intron/exon boundary and is not predicted to result in any splicing changes.

11.4 Discussion

Using a combination of exome sequencing and autozygosity mapping, followed by appropriate variant filtration, we set out to establish the genetic cause of disease in a consanguineous Pakistani kindred, in which affected individuals exhibited infantile-onset, severe choreodystonia and cerebellar signs. We identified 4 potentially causal variants. Of the genes affected, *TRIM3* seemed the most biologically plausible as it was highly expressed in the brain and, more precisely, in the cerebellum. Moreover, it has functional similarities to the protein mutated in the *dt* rat model of dystonia, caytaxin, in that both interact with memebers of the kinesin family of proteins to mediate transport of specific cargos. Furthermore, experimental evidence has demonstrated defects in trafficking of $GABA_A$ receptors in TRIM3 knockdown models, whilst defective GABAergic transmission in Purkinje fibres is now known to be the pathogenic mechanism underlying the dystonic phenotype in *dt* rats.

On a genetic level, however, the variant detected in this gene is not particularly convincing. Although located next door to a critical, ultra-conserved histidine residue that is involved in zinc ion coordination, the glycine to alanine substitute is universally predicted to be benign (both are small non-polar amino acids) and interspecies conservation is not complete at this position. In addition, we could find no other potentially causal mutations in a cohort of generalised dystonia cases (though this in itself is not particularly surprising given the rarity of autosomal recessive dystonia and the difficulty in matching the phenotype in a sizable cohort).

In summary, at present it is not possible to say with any certainty what the genetic cause of the dystonia in this family is. A number of different possibile explanations exist for this. These are:

1) The causal variant is located in the 10-15% of the CCDS coding bases within the homozygous regions that were not covered by exome sequencing.
2) The genetic cause of this disease is located outside of the CCDS coding region, i.e. it is intronic, intragenic or located within a non-coding RNA.
3) The genetic cause of this disease is an exonic copy number variant that would not be expected to be detected by either exome sequencing or by use of genome wide SNP data.
4) Either *ZNF195*, *OR52D1* or *DCHS1* is in fact the cause of the disease and we have pursued the incorrect variant.
5) *TRIM3* is the cause of the disease, but we have not found a second mutation as phenotypic matching in the screening cohort was suboptimal or the disease is simply too rare to expect to find a second case in a cohort of this size.

With respect to explanations 1 and 2, whole genome sequencing may hold out hope that these problems could be overcome in the near future. However, even with 'whole genome' sequencing there are likely to be areas that remain uncovered due to the physical features of the DNA, such as repetitiveness and GC content. Moreover, whole genome sequencing comes with it own set of problems, in particular the current lack of any large-scale databases of whole-genome variation in normal individuals (particularly those of non-white ethnicity) and the absence of any currently validated means of predicting the likely pathogenicity of non-coding variants.

With respect to explanation 5, although we attempted to maximize our chances of screening autosomal recessive cases by selecting samples from younger individuals with either no family history or only affected siblings, it is inevitable that a proportion of the cases screened will, in fact, have been either non-genetic or, if genetic, due either to genes associated with autosomal dominant inheritance (because of incorrectly recorded family history, *de novo* mutations, or reduced penetrance, for example).

CHAPTER 12:

An Association Study In Neuropathologically Proven PD

12. An Association Study In Neuropathologically Proven PD

12.1 Introduction

To date, several large-scale genome-wide association studies (GWAS) have been completed for Parkinson's disease (PD), culminating in a recently published meta-analysis that identified 16 loci surpassing the threshold for genome-wide significance [3, 4, 58, 448, 449]. Of these 16 loci, some are predictable findings, whilst others offer the new insights into the pathogenesis of the disease. For example, from a neuropathological perspective, the discovery that variants in SNCA act as a risk factor for PD is consistent with the deposition of aggregated α-synuclein in the PD brain. However, the detection of variation in MAPT as a risk factor in PD was unexpected, as the relevance of tau aggregation in PD remains unclear.

Predictably, a number of tauopathies are associated with common or rare variants in MAPT, including corticobasal degeneration, frontotemporal dementia with parkinsonism linked to chromosome 17, and progressive supranuclear palsy (PSP) [450]. In fact, variation in MAPT is the strongest genetic risk factor for PSP. The association with this condition is driven by a series of polymorphisms in near complete linkage disequilibrium, which form an extended haplotype, termed H1, that covers the entire gene [451]. Inheritance of two copies of the risk haplotype in this region confers an odds ratio for developing PSP of ~4 [451].

It is a well known fact that there can be considerable clinical overlap between PD and PSP, especially in the earlier stages of the diseases. Even with increased awareness, the rate of misdiagnosis of idiopathic PD has been estimated at between 10–25% in autopsy studies, and 6–26% in community studies [452, 453]. In one autopsy study of 100 cases of clinically diagnosed PD, 24 cases were found to be misdiagnosed and 6 of these were PSP [452]. It is probable, therefore, that large GWAS, which rely on clinically diagnosed PD cases, will contain some individuals with PSP. Despite the potential for clinical confusion, PSP is neuropathologically distinct from idiopathic PD. PSP is classified as a primary tauopathy and is characterised morphologically by deposition of

four-repeat tau in neurons as neurofibrillary tangles and in both astrocytes and oligodendroglia as tufted astrocytes and coiled bodies, respectively. The morphological characteristics and anatomical distribution of the tau pathology and the biochemical composition of the tau lesions are different from those seen in Alzheimer's disease [454]. PD, on the other hand, is classified as a synucleinopathy, characterised by abnormal fibrillar cytoplasmic inclusions, termed Lewy bodies, of which the principal protein component is α-synuclein [455].

Given the strength of the association between MAPT with PSP, it has been suggested that the tau signal seen in Parkinson's disease association studies might be the result of unrecognised PSP contamination of the case cohort. Therefore, in order to answer this important question, we performed a focused analysis of our pathologically proven PD cases using genotyping data from our own in-house GWAS datasets [3] as well as newly genotyped PD brains specifically for this study.

12.2 Subjects, Materials and Methods

12.2.1 Sourcing of Tissue Resources

Cases were selected from two brain banks in London with a confirmed primary neuropathological diagnosis of PD. Neuropathological diagnosis had been made by an experienced neuropathologist and was based on accepted morphological criteria [456]. Cases were excluded who exhibited a family history of Parkinson's disease consistent with Mendelian inheritance of the disease.

12.2.2 Extraction of DNA from Brain Tissue

DNA from brain tissue was extracted as per section 4.2.3

12.2.3 Genotyping of New Cases

Newly extracted DNA was genotyped as per section 4.10 using the Immunochip, a custom SNP chip supplied by Illumina. Although absolute numbers of SNPs genotyped by the Immunochip is relatively modest (appox. 200,000 SNPs), the strategy employed in its design meant that the loci likely to be associated with PD were disproportionately well covered. Genotyping data was viewed in Genome Studio for

initial quality control and, subsequently, PLINK files generated using a publically available plug-in.

12.2.4 Selection of Previously Genotyped Cases

In total, 469 DNA samples from the UK included in the original GWAS discovery and replication phases were thought to be derived from neuropathologically proven PD cases. During this work, however, it became apparent that tissue had often been sent on the basis of clinical diagnosis alone and, unfortunately, no effort had been made to check the neuropathologically diagnosis was in fact iPD.

We therefore undertook a manual review of the clinical history and neuropathological diagnosis for all of these cases. Inclusion criteria were a neuropathological diagnosis consistent with iPD and an age at onset of greater than 40 years of age. Exclusion criteria were an alternative primary neuropathogical diagnosis, a family history of PD consistent with Mendelian inheritance of the disease, or an age at onset of less than 40 years of age.

In total 46 were excluded because they did not meet the above criteria. This included 12 cases of PSP. 423 remained for inclusion in the analysis. As above, genotyping data was viewed in Genome Studio for initial quality control and, subsequently, PLINK files generated using a publically available plug-in.

12.2.5 Generation and Quality Control of Control Genotyping Data

Genotyping data for 5,200 control individuals from the UK had been obtained using the Quad 660 SNP chip for the discovery phase of the GWAS. For the subsequent replication stage of the study, 4,537 control individuals from the UK were genotyped using the Immunochip custom SNP chip. Unfortunately, it was known that there was likely to be some overlap between the two populations, which might artificially inflate the results of any meta-analysis. As a first step, therefore, we sought to remove definite duplicates from the control population and individuals who are likely to be siblings.

Treating both datasets separately initially, PLINK files were generated from the raw genotyping data using Genome Studio. For each dataset, SNPs were eliminated that showed a genotyping rate of less than 90% across the dataset or were not in Hardy-Weinberg equilibrium. Next, for each dataset in turn, individuals with a genotyping failure rate of greater than 10% were also removed. Finally, both datasets were merged and SNPs were eliminated with a genotyping rate of less than 90% across the entirety of newly merged dataset (which essentially had the effect of leaving only high quality SNPs that were common to both genotyping platforms for consideration). In the end, genotyping information was available for all 9737 individuals across 20,485 markers.

Pairwise identity by descent (IBD) was calculated for this merged dataset using PLINK. Each pairwise comparison results in a measure (PI-HAT), ranging from 0 to 1, that represents the proportion of the genome that is likely to be identical by descent. In order to accurately estimate IBD, it is desirable to consider greater than 100,000 independent SNPs, with higher numbers leading to increasing accuracy. Under these optimal conditions, a PI-HAT close to 1 represents a duplicated sample, a PI-HAT of around 0.5 represents a sibling, parent or child, a PI-HAT of around 0.25 represents a cousin. However, for the purposes this study, we were limited to the 20,485 markers shared between both platforms. It was recognised that this would lead to a some inflation in the value of PI-HAT due to the reduction in the number of independent markers which PLINK would be able to consider. We therefore chose an arbitrary cut-off point of a PI-HAT of greater than 0.3, which we felt would capture all duplicate samples, all 1st degree relatives and most if not all 1st cousins. In practice, this identified 2608 pairs of samples that were likely to be related at this level and one member of each pair was eliminated from the original unmerged datasets, taking care to remove roughly even numbers from each.

12.2.6 Identification of Outliers and Matching of Cases to Controls

Accurate detection of population stratification requires genome-wide genotyping data. Since only approximately 20,000 SNPs were shared between the two platforms used for genotyping in this study, it was necessary to treat each set of samples separately.

Information regarding identity by state was generated using PLINK for both datasets in turn for use in the following steps, which were performed on each dataset in turn.

As a first step, aimed at matching the controls as closely as possible to the cases in terms of population of origin, clustering was performed on the basis of identity by state (IBS) using PLINK in order to sort samples into groups by genetic proximity. Each group was permitted to hold a maximum of 1 case and up to 10 controls. Groups were then reviewed to find those groups than did not contain any case (suggesting that the controls contained therein were significantly distinct from the case population) and all controls contained within such groups were eliminated. This resulted in the elimination of 1352 controls from the GWAS discovery dataset and 846 controls from the GWAS replication dataset.

As a second step, to detect outliers in both cases and controls, each sample in turn was paired with its closest neighbour in terms of IBS. We then determined whether the proband's closest neighbour was significantly more distant to the proband than all other individuals' nearest neighbour is to them. In other words, from the distribution of 'nearest neighbour' scores, one for each individual, we calculated using PLINK a sample mean and variance and transformed this measure into a Z score. Individuals with a Z score of -1 in relation to their nearest neighbour (i.e. one standard deviation of difference) were eliminated. The process removed a further 58 individuals for removal from the GWAS discovery dataset and a further 130 individuals from the GWAS replication dataset.

12.2.7 Selection of SNPs for Testing

The primary analysis focused on a region approximately 500 kilobase (kb) either side of *MAPT*. The single nucleotide polymorphisms (SNPs) tested for association in this region were selected on the basis that they had achieved genome-wide significance in the UK samples used in the discovery phase of our GWAS meta-analysis and had also been genotyped on the custom-built immunochip used for the replication phase. Eight SNPs met these requirements (rs393152, rs1635291, rs7215239, rs1237319, rs17690703, rs17769552, rs1981997, and rs8070723). None of the SNPs in *MAPT*

chosen by this method are in significant linkage disequilibrium with the SNPs used by Vandrocova et al. (2009) in their *MAPT* haplotype analysis (maximum $r2 \approx 0.2$).

In earlier GWAS, the region surrounding α-synuclein (SNCA) has consistently been shown to be the locus most strongly associated with PD (International Parkinson's Disease Genomics Consortium et al., 2011; UK Parkinson's Disease Consortium et al., 2011; Satake et al., 2009). Thus, in order to verify that this analysis had sufficient power to detect previously well-documented associations, we also included the SNPs in a second 1 megabase (Mb) region spanning SNCA that met the same requirements as above. Two SNPs were available for testing (rs356220 and rs2736990).

12.2.8 Association Analyses and Meta-Analysis

After implementation of the measure described above, there was genotyping data for 301 cases and 2455 controls obtained using the Quad 610 SNP chip and 289 cases and 2301 controls using the Immunochip custom-built SNP chip. Association between genotype at each of the 10 markers indicated above and the subject phenotype (case/control) was determined for each dataset independently. The association analysis was performed using PLINK. Although the minimum minor allele frequency (MAF) was set to 5% as is considered best practice, in reality the method of choosing SNPs for testing in this study meant that all SNPs would necessarily have a MAF of greater than 5%. Subsequently, a meta-analysis of the 2 association analyses was performed using the same software. Fixed and random effects models were used to generate *p*-values and the I^2 index was used to quantify heterogeneity between the studies.

12.3 Results

The primary analysis demonstrated statistically significant association in 8 SNPs spanning the *MAPT* gene (see table 35 and figure 56). The SNP showing strongest association in this study was rs17690703, which had a p-value of 8.9×10^{-5} and was calculated to confer an odds ratio of 0.76 per minor allele dose.

Table 35 - SNPs in a 1Mb region spanning MAPT showing statistically significant association with p-values in fixed and random effects models, I^2 index, minor allele frequency (MAF) and odds ratio per minor allele dose.

SNP reference	Physical Position	p-value (fixed effects)	p-value (random effects)	I2 index	MAF	Odds ratio per minor allele dose
rs393152	41074926	2.0 x 10-4	2.0 x 10-4	0.00	0.24	0.76
rs1635291	41107696	5.0 x 10-4	5.0 x 10-4	0.00	0.26	0.77
rs7215239	41123556	3.2 x 10-4	3.2 x 10-4	0.00	0.26	0.76
rs12373139	41279910	2.1 x 10-4	2.1 x 10-4	0.00	0.24	0.76
rs17690703	41281077	8.9 x 10-5	8.9 x 10-5	0.00	0.28	0.75
rs17769552	41283070	9.9 x 10-5	9.9 x 10-5	0.00	0.24	0.74
rs1981997	41412603	2.7 x 10-4	2.7 x 10-4	0.00	0.24	0.76
rs8070723	41436901	1.8 x 10-4	1.8 x 10-4	0.00	0.24	0.76

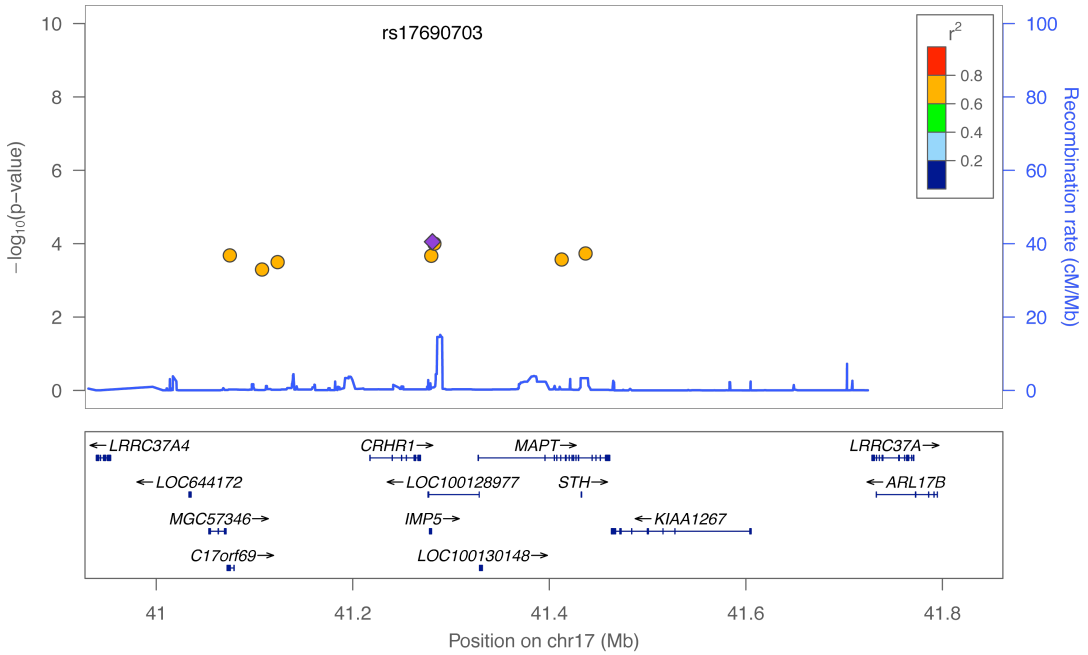

Figure 56 – Physical position of the SNPs tested across the MAPT locus with p-values, recombination rate and linkage disequilibrium (LD). A diamond represents the top hit, which is the reference for the LD calculations.

The results of the secondary analysis using SNPs spanning *SNCA* confirms this study's ability to detect the predicted association with this region using only the relatively small sub-population of neuropathologically confirmed cases of PD (see table 36 and figure 57). rs356220, the SNP showing the strongest association in this study, had a p-value of 5.4×10^{-4} was calculated to confer an odds ratio of 1.24 per minor allele dose. Furthermore, this SNP is in high linkage disequilibrium ($r^2>0.9$) with and shows the same direction of effect as the top SNP from the largest meta-analysis published to date (rs356219; not included in this study as not common to both genotyping platforms) suggesting a consistent pattern of association.

Table 36 - SNPs in a 1Mb region spanning SNCA showing statistically significant association with p-values in fixed and random effects models, I^2 index, minor allele frequency (MAF) and odds ratio per minor allele dose.

SNP reference	Physical Position	p-value (fixed effects)	p-value (random effects)	I2 index	MAF	Odds ratio per minor allele dose
rs356220	90860363	5.4×10^{-4}	5.4×10^{-4}	0.00	0.36	1.24
rs2736990	90897564	6.2×10^{-3}	6.2×10^{-3}	0.00	0.45	1.18

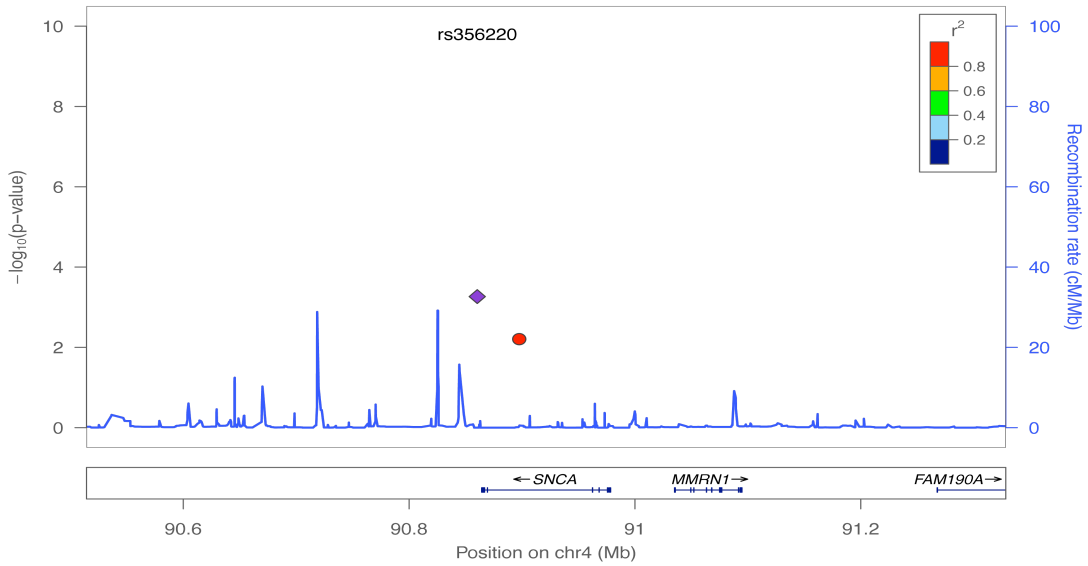

Figure 57 - Physical position of the SNPs tested across the SNCA locus with p-values, recombination rate and linkage disequilibrium (LD). A diamond represents the top hit, which is the reference for the LD calculations.

12.4 Discussion

Clinical misdiagnosis has the potential to introduce contamination into association analysis. With regard to PD GWAS, PSP contamination is of particular concern as the association between PSP and the MAPT locus is particularly strong and studies suggest clinical misdiagnosis of PSP as PD is relatively common. Therefore, in this analysis, we relied solely on neuropathological diagnosis as the inclusion criterion for the case cohort.

The results confirm that the region surrounding *MAPT* is associated with idiopathic PD in our neuropathologically proven PD cohort and, therefore, that this association does not result from contamination with primary tauopathies, such as PSP. Association at the SNCA locus was also seen, confirming the analysis's power to detect known associations even with these moderate sample sizes. Although the gene most likely to be responsible for the association signal remains *MAPT* itself, this gene is located in a block of near complete linkage disequilibrium that extends over a total of nearly 2Mb on chromosome 17. It is conceivable, therefore, that a different gene within this haplotype block may in fact be driving the association.

Accepting *MAPT* as the most likely candidate, the persistence of the association at this locus in pathologically proven PD raises the possibility that that dysfunction of tau may in fact be pathogenic in PD. Variants in MAPT may act either to increase expression or alter the splicing of tau so as to favour its aggregation. It is already known that the H1c haplotype, which underlies the risk in PSP, results in increased expression of tau, particularly of four repeat containing transcripts [457]. Recently, it has also become increasingly clear that, in many neurodegenerative diseases, the aggregation of one protein may often be associated with or even induce the aggregation of others [458]. On a genetic level, studies in clinically defined PD have suggested that genotypes at the MAPT and SNCA loci act synergistically to confer susceptibility to PD and that variation at MAPT may be particularly associated with cognitive decline [459, 460]. Certainly, Alzheimer-like tau pathology has long been known to co-exist with the morphological changes typical of idiopathic PD, where it may be associated with cognitive dysfunction [461]. This has mostly been viewed as a consequence of pathological

ageing or co-existent early Alzheimer's disease, rather than as being an intrinsic part of the pathogenic pathway of PD itself. Co-existent Alzheimer's disease seems, however, unlikely to be able to fully explain the MAPT association in idiopathic PD as a recent large GWAS in Alzheimer's disease failed to demonstrate any association at this locus [462]. Significant tau pathology has also been noted in some monogenic forms of PD, including in some cases resulting from *LRRK2* mutation and in some members of the Contursi kindred, who carry the A53T *SNCA* mutation, as well as in α-synuclein overexpressing transgenic mice [463, 464]. More recently, it was found that the distribution of tauopathy in the brains of individuals with PD was distinct from that observed in Alzheimer's disease, with hyperphosphorylated tau evident in striatum, but not in the inferior frontal gyrus despite an increase in α-synuclein in this region [465].

It is likely that more detailed neuropathological studies of tau pathology in the brains of individuals with idiopathic and genetic forms of PD will be required to fully dissect out its role in the pathogenesis of this condition. One promising approach may be to look for a correlation between the extent of tau pathology in idiopathic PD and the genotype at the *MAPT* locus. The existence of such a correlation would support the idea of a direct effect of these variants on tau aggregation.

CHAPTER 13:

Mutational Screening of VPS35 in a UK Parkinson's Disease Cohort

13. Mutational Screening of VPS35 in a UK Parkinson's Disease Cohort

13.1 Introduction

In 2011, using an whole-exome sequencing approach, two independent groups identified a missense mutation in vacuolar protein sorting 35 homolog (*VPS35* c.1858G>A; p.Asp620Asn) as the probable cause of late-onset PD in a number of kindreds.[466] Pathogenicity was supported by segregation, evolutionary conservation at that base, and software predictions that the variant is likely to be damaging. However, no other pathogenic mutations had been identified with certainty in *VPS35*, and the c1858G>A mutation had only ever been found in Caucasian individuals. Definitive evidence that a gene such as VPS35 predisposes to familial Parkinson's disease usually requires the identification of other mutations which segregate with Parkinson's disease in families, and the observation that *VPS35* mutations occur in PD patients in other populations. Moreover, the at the time of publication, it was not known what proportion of unexplained familial Parkinsonism VPS35 mutations might account for.

In order to attempt to answer these important questions, we screened our own Parkinson's population for variants in this gene, both in order to estimate the frequency of the published mutations and in order to search for novel mutations that may be disease-causing.

13.2 Subjects, Materials and Methods

13.2.1 Selection of Cases of Mutational Screening

The study included 160 familial PD cases, 175 young-onset PD cases, and 262 sporadic, neuropathologically confirmed PD cases (total number screened, 501). Neuropathologically confirmed cases were selected from the Queen Square and (Parkinson's disease United Kingdom) brain bank. The study was approved by the East Central London Research Ethics Committee 1 and informed consent was obtained as per its guidelines.

All living patients fulfilled criteria for clinical diagnosis of PD at the time of the study with at least 2 of 3 cardinal signs of tremor, rigidity, and bradykinesia, as well as a positive response to levodopa therapy. Familial cases were defined as those reporting 1 or more first degree relatives with PD with a pedigree consistent with autosomal dominant inheritance and negative testing for known mutations in *SNCA* and *LRRK2*. Young-onset PD cases were defined as those who had developed their first signs of the disease at age 45 or younger (mean age of onset in young-onset PD was 37 ± 6, range 14–45 with a male to female ratio of 1.63:1). A family history of PD (first or second degree relative) was noted in 31 young-onset PD cases (17.7%). The neuropathological diagnosis for brain bank cases was made by an experienced neuropathologist and based on accepted morphological criteria (Ince et al., 2008). This group included 166 males and 96 females, with an average age at onset of 64 ± 11 years (range, 29–85) and a family history recorded in 7.6% of cases.

13.2.2 Mutational Screening

VPS35 exons 2–12 are located within a region that is duplicated 12 megabases (Mb) upstream. Primer pairs were designed specifically to amplify these exons using Gene Runner (v.3.0.5, Hastings Software Inc., Hastings, NY, USA) using Ensembl reference transcript ENST00000299138. We amplified all exons and exon-intron boundaries using the following PCR reaction mix: 10ng of genomic DNA, 10nM forward primer, 10nM reverse primer and 10ul of FastStart PCR Master (Roche, IN, USA). To amplify exon 9 and 10, we added 1ul of 5% dimethylsulfoxide (DMSO; American Bioanalytical, MA, and USA) to the PCR mix. Sequencing was accomplished as per sections 4.6 – 4.11.

All 17 exons and exon-intron boundaries of VPS35 were sequenced in 96 familial PD cases, and exon 15 (in which the c.1858G>A; p.Asp620Asn mutation is found) in an additional 64 familial PD cases, 175 young-onset PD cases, and 262 sporadic, neuropathologically confirmed PD cases (total number screened, 501).

13.2.3 Clinical Characterisation of Individual's Testing Positive of a VPS35 Variant

Clinical details were collated were possible from living family members for those affected relatives that were deceased. Where possible affected individuals were clinically examined and olfactory testing undertaken using UPSIT.[467]

13.3 Results

13.3.1 Results of Mutational Screening

Screening revealed a single case from our familial PD cohort who harbored the previously described p.Asp620Asn mutation. No other nonsynonymous variants were detected in any other exon of any sample. In addition to this, we observed 7 other sequence variants, including 2 novel and 5 previously described variants (Table 37 overleaf

13.3.2 Clinical Characterisation of Kindred with p.Asp620Asn Mutation

In order to determine whether the p.Asp620Asn mutation segregated with PD in the family of the individual from our FPD cohort, we contacted other affected family members: a sister, and two cousins, they subsequently were found also to have the p.Asp620Asn mutation. The family is of European ancestry, the pedigree (figure 58) is consistent with autosomal dominant mode of inheritance and revealed 9 further affected individuals across 3 generations, all of whom are deceased.

The index case (IV-7) noticed his first symptoms in 1993 at the age of 40, consisting of stiffness in the left arm on waking followed, a few months later, by gradually worsening stiffness in the left leg. He was diagnosed as suffering from PD in 1995. At that time, examination showed unilateral bradykinesia and reduced arm swing. L-dopa was commenced with a good response, although he continued to gradually worsen. He began to notice a mild intermittent tremor of the left upper limb 13 years after his initial symptoms, followed by the spread of the stiffness to his right upper limb.

Table 37 – Details of all variants detected in *VPS35* by mutational screening.

Nature of Variant	Nucleotide change	Amino acid change	Rs number (if available)	Exon	Number of case
Non-synonymous	c.1858G>A	p.Asp620Asn	Recently published	15	1
Synonymous	c.231T>C	p.Leu77Leu	Rs11550462	4	1
	c.1842T>C	p.Tyr615Tyr	Novel	15	1
	c.1938C>T	p.His646His	Rs16875	15	1
UTR	c.1-34G>A		Rs3743928		1
Intronic	c.3+25A>C		Novel	1	1
	c.1524+42G>C		Rs4966616	12	27 (2 homo)
	c.1648+26G>C		Rs2304492	14	58 (10 homo)

Dyskinesias started after 9 years of L-dopa therapy, but were not severe. He complains of difficulty with balance, but has not had any falls. There is no history of mood disturbance or other psychiatric features, including visual hallucinations. Cognitively, he feels that his concentration has deteriorated somewhat over the years, but he has only recently retired from a demanding job. A Mini Mental State Examination (MMSE) performed 18 years into the disease revealed a score of 28/30. Olfactory testing was undertaken 18 years into the disease, using the 40-item University of Pennsylvania Smell Identification Test (UPSIT) [467] and results were interpreted in comparison with 55 control subjects (27 males and 28 females) and 46 PD patients (30 males and 16 females) who had been previously tested for a study about olfaction in the UK[468]. The patient correctly identified 29/40 items of the UPSIT, scoring on the 37th percentile of 55 controls aged between 55 and 65 from the UK, and the 87th percentile of 46 PD subjects aged between 55 and 65 from the UK. His sibling (IV-8) developed symptoms aged 52 years of age and was formally diagnosed 2 years later. Initial symptoms included cramping in the left foot and tripping when walking. Within 6 months she developed a unilateral leg tremor. Further symptoms included micrographia, and freezing.

Dopamine agonists were tried initially but within 3 years she required L-dopa. She has not developed any dyskinesias. She does not report any cognitive symptoms and a Mini Mental State Examination was 30/30. Olfactory testing was undertaken 11 years into the disease as described above with the patient also identifying 29/40 items of the UPSIT. Individual IV-5 presented with a stiffness and rigidity of the left leg at the age, followed by walking difficulties, bradykinesia, loss of postural reflexes and micrographia aged 47. There was no rest tremor. He was diagnosed shortly afterward. He responded well to L-dopa, but developed dyskinesias after 10 to 15 years of treatment and has now had a deep brain stimulator implanted to good effect. There have been no hallucinations or psychiatric complications. The clinical history for individual IV-6, was provided by her brother. She developed PD at the age of 41 with noticeable tremor and generalized bradykinesia and rigidity. She responded well to L-dopa, developing only mild dyskinesias after a reasonable period of treatment. She underwent a pallidotomy at around 60 years of age and died at age 66 as a consequence of a nosocomial infection after a hip replacement operation.

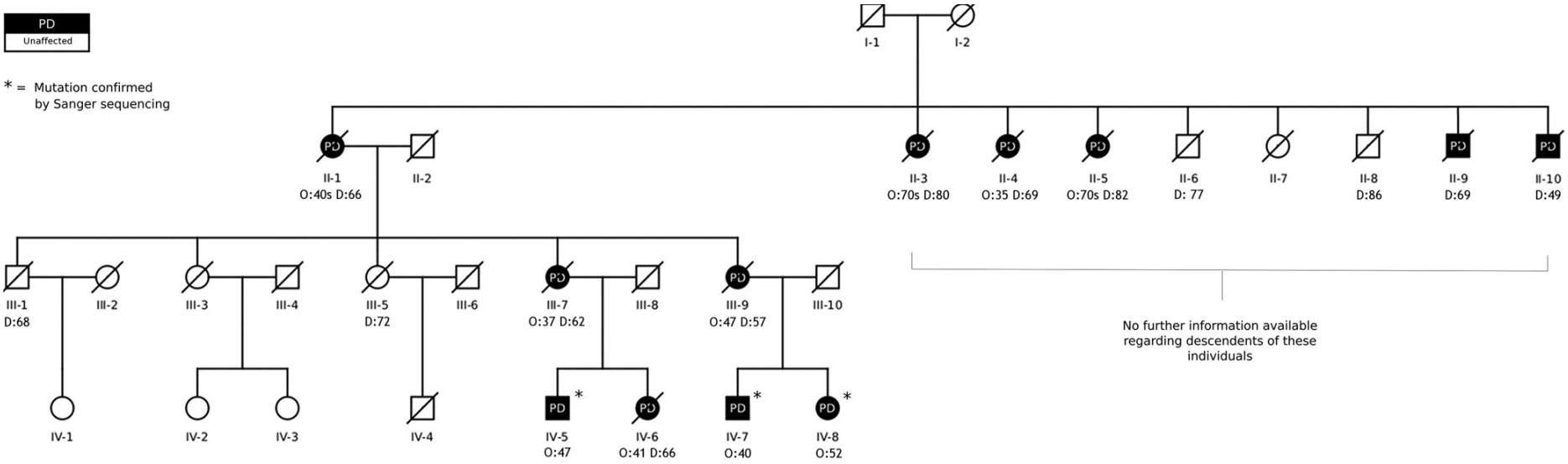

Figure 58 – Pedigree of a family showing highly penetrant autosomal dominant inheritance of Parkinson's disease (PD). Age of onset (O:) and age at death (D:) are indicated where known for all descendants of I-1 and I-2. The p.Asp620Asn mutation of *VPS35* was confirmed by Sanger sequencing in the 3 living individuals affected by the disease who were available for testing.

Limited information is available for other affected family members. Individual II-1 presented with akinetic-rigid parkinsonism in her late fourth decade, family members noted that she never developed tremor, but suffered with depression late in the disease course. She died aged 66 years. Individuals II-3 and II-5 developed parkinsonism in their seventh decade and died in their eighth decade. II-4 developed tremor-predominant Parkinson's disease aged 35 years and had a slowly progressive disease course, dying at the age of 69. Individuals II-9 and II-10 are known to have been affected with parkinsonism, but ages of onset are not known. They died in their sixth and fourth decade respectively. Individual III-7 developed parkinsonism at age 37. Medical records and family members indicate left-sided rigidity, bradykinesia, hypomimia, falls, and freezing. Tremor was not recorded. Surgical interventions, including pallidotomy (right then left) and thalamotomy, were performed before L-dopa was commenced. No dyskinesias developed despite 10 years of treatment and she died at age 62. Individual III-9 developed parkinsonism aged 47 years presenting with a unilateral rest tremor. Family members report early severe muscular spasms, difficulty walking, and falls. Later in the disease, she developed depression and cognitive impairment. She died aged 57 years.

13.4 Discussion

Using exome sequencing, two groups independently identified a single c.1858G>A mutation in *VPS35* as the probable cause of Parkinson's disease in several kindreds. Therefore, in this study, we screened our own Parkinson's population for variants in this gene, both in order to estimate the frequency of the published mutations and in order to search for novel mutations that may be disease-causing. We included PD patients with a positive family history, as well as those with early onset and late-onset sporadic disease in order to maximize our chances of identifying mutations.

We identified 1 individual with the published putative disease-causing mutation (c.1858G>A; p.Asp620Asn). This individual had a family history that was consistent with highly penetrant, autosomal dominant Parkinson's disease. Based on clinical examination, notes and family reports, the disease in this family appears clinically similar to idiopathic PD. Onset was generally unilateral, with slow progression and

variable tremor. Cognitive or psychiatric features are not prominent. Response to L-dopa was generally good and was not associated with early or unusually severe dyskinesias. The 2 patients in which olfaction was tested here show mild to moderate olfactory dysfunction when compared with controls in the UK, but still performed better than the vast majority of PD patients. Olfaction has been demonstrated to be unimpaired in homozygous and compound heterozygous carriers of Parkin mutations [469, 470] but impaired in *LRRK2* parkinsonian carriers [471-474], although not to the same extent as in sporadic PD. The most notable feature of our kindred is the relatively early age at onset. Six individuals developed PD in their third or fourth decades, of which 4 were younger than the age of 45.

No other potentially disease-causing mutations were found in exon 15 (597 cases screened) or in any other exon (96 cases screened). It would seem reasonable to conclude, therefore, that the recently published *VPS35* c.1858G>A mutation, is not a common cause of PD, at least in our cohort of patients.

CHAPTER 14:

Mutational Screening of GNAL in a UK Cervical Dystonia Cohort

14. Mutational Screening of GNAL in a UK Cervical Dystonia Cohort

14.1 Introduction

In 2013, using an whole-exome sequencing approach, mutations in *GNAL* were identified as a novel cause of primary dystonia by one group [171] and confirmed shortly afterwards by another [364]. Mutations in this gene appear to cause autosomal dominant, primary dystonia with a cervical predilection and evidence of incomplete penetrance [171].

At the time of this work, the frequency of mutations in *GNAL* as a cause of dystonia remains unclear. In the initial discovery paper by Fuchs *et al.*, researchers reported *GNAL* mutations in 6 out of 39 families screened (~15%), whereas, in the subsequent study by Vemula *et al.*, only 3 *GNAL* mutations were detected in 760 subjects with familial or sporadic primary dystonia (<0.5%). Although the difference in choice of cohort to screen makes direct comparison difficult, there was clearly a larger than expected difference between the two studies, which made it difficult to judge the true frequency of *GNAL* mutations in primary dystonia. Yet, this information is clinically relevant, as a mutation frequency of ~15% would justify early and widespread genetic testing of *GNAL* in familial dystonia and probably also in younger onset cervical dystonia; a frequency of <0.5%, on the other hand, would not.

In order to attempt to answer the question of just how prevalent GNAL mutations might be as a cause of familial or, indeed, apparently sporadic cervical dystonia, we performed mutational screening of the gene by Sanger sequencing in 192 probands with either familial of sporadic cervical dystonia.

14.2 Subjects, Materials and Methods

14.2.1 Selection of Cases of Mutational Screening

Participants' DNA was selected from a library of samples donated with prior consent for research use on the basis of a clinical description of focal or segmental dystonia that included the cervical region. 136 samples belonged to females and 56 to males. A family history, defined as one or more first or second-degree relatives with dystonia, was

recorded in 84 cases. All familial cases had been screened for *TOR1A* and *THAP1* mutations and were negative. The study was approved by the East Central London Research Ethics Committee 1 and informed consent was obtained as per its guidelines.

14.2.2 Mutational Screening

Two isoforms of *GNAL* are expressed in the brain: isoform 1 (ENST00000334049) is the longest, whereas isoform 2 (ENST00000423027) is the major isoform. Primers were thus designed to amplify all exons, the exon/intron boundaries and the 5'UTRs of both these isoforms. Sequencing was accomplished as per sections 4.6 – 4.11.

14.2.3 Segregation Analysis of Kindreds with GNAL Variants

Clinical details were collated found to habour *GNAL* variants. A family history was established and the wider family was invited to contact us to provide a history and, where possible, attend for examination. DNA was obtained from all available members of the wider kindred – whether affected or unaffected – and the relevant exon of GNAL was screened to establish mutational status within the family.

14.3 Results

14.3.1 Mutational Screening and Segregation Analysis

We identified only two novel single nucleotide variations in *GNAL* in three individuals in our case cohort. Neither variant was annotated in dbSNP, 1000 genomes, the NHLBI Exome Sequencing Project or cg69 databases. The first was a missense mutation in exon 2 of the gene (cDNA.1053C>T; P149S in ENST00000334049) that was detected in one individual with onset of cervical dystonia in the 4th decade. It was predicted to be benign by SIFT and PolyPhen2, but disease-causing by Mutation Taster (albeit with a low probability of 0.72). The individual concerned exhibited early onset (3rd decade) cervical dystonia with a similarly affected father, suggesting autosomal dominant inheritance. However, segregation analysis revealed the variant had in fact been inherited from his unaffected mother and was also present in his unaffected brother (see figure 59A), ruling it out as the cause of the dystonia in this family.

The second variant, located in the 5'UTR of isoform 2 (cDNA.199C>T in ENST00000423027), was found in two individuals in our cohort. Both had onset of cervical dystonia in the 6th decade of life, but one individual was a sporadic case, whilst the other was part of family with multiple affected members (see figure 59B and 59C). The variant affected a highly conserved base (PhyloP score = 4.158) and was predicted to be disease causing by MutationTaster. SIFT and PolyPhen2 are unable to offer predictions on non-coding variants.

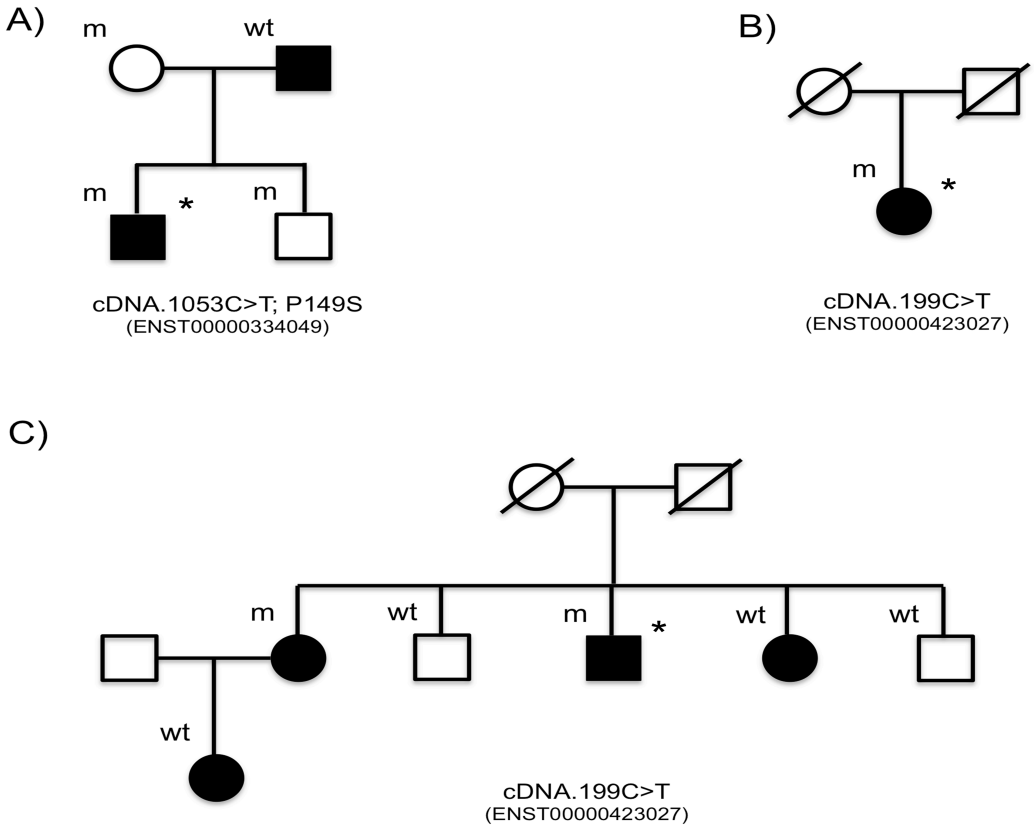

Figure 59 – Genetic pedigrees and results of segregation analysis of kindreds with variants detected in *GNAL*. Affected individuals are indicated by solid symbols. None of the detected variants segregate with disease, suggesting they are not pathogenic. m = heterozygous carrier of the variant indicated below each kindred; wt = homozygous wildtype alleles

14.4 Discussion

In conclusion, this work did not identify any mutations in *GNAL* that could be cause of the dystonia in 196 cases drawn from the UK, including 84 familial cases. Data presented herein suggest that *GNAL* mutations do not represent a common cause of dystonia – in the UK population at least – and that the overall frequency of *GNAL* mutations may be closer to the figure obtained by Vemula *et al.* [364] than the 15% initially reported by Fuchs *et al.* [171]. This study also emphasises the importance of segregation studies in establishing the pathogenicity or otherwise of novel variants.

CHAPTER 15:

The Role of EIF4G1 as a Parkinson' Disease Gene

15. The Role of EIF4G1 as a Parkinson's Disease Gene

15.1 Introduction

In September 2011, Chartier-Harlin and colleagues reported that they had determined the genetic cause of the Parkinson's disease in a multi-incident, autosomal dominant kindred from Northern France (see figure 60 below) [475]. Using a combination of linkage analysis and mutational screening of the coding exons of all 159 genes within the region of best linkage, they identified 5 novel variants that were present in 3 affected family members. Subsequent segregation analysis in all 10 blood-related, affected individuals demonstrated that only one variant (c.3614G>A; p.R1205H) in exon 24 of *EIF4G1* fully segregated with the disease.

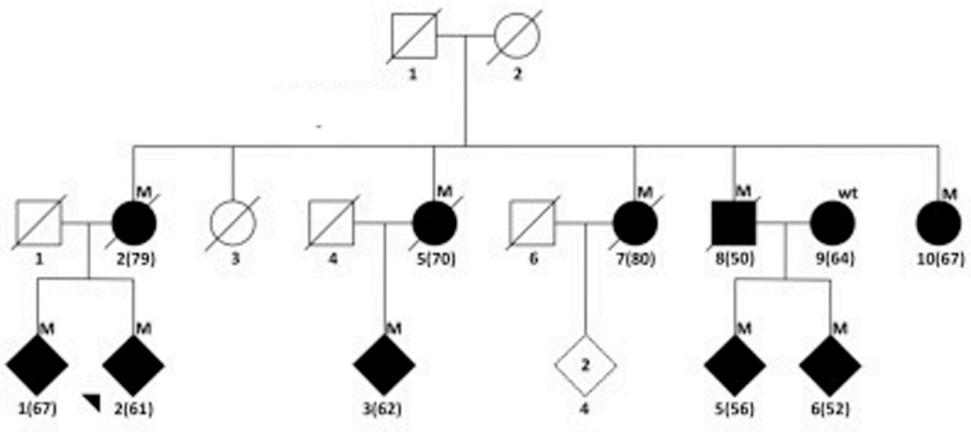

Figure 60 - Structure of index family as reported in Chartier-Harlin et al. (2011). The ages of onset for affected individuals are shown in brackets. An 'M' above the genetic symbol indicates that the individual was heterozygous for the EIF4G1 c.3614G>A mutation; 'wt' indicates the affected individual was homozygous for the normal allele.

Subsequent screening of 4708 individuals with idiopathic PD identified a further 7 probands who were heterozygous for this variant, though only 3 individuals showed any evidence of familial inheritance of disease. More than one affected individual was available for segregation analysis in only 2 these familial and in both cases the affected individuals were siblings making the probability of segregation by chance alone high. Conversely, the variant was not detected at all in 4050 control subjects.

Mutational screening of the full 31 coding exons of the *EIF4G1* gene in 225 probands with PD and 185 ethnically matched controls identified identified a further four four missense variants not observed in controls, that is c.1505C>T; p.Ala502Val, c.2056G>T; p.Gly686Cys, c.3490A>C; p.Ser1164Arg, and c.3589C>T; p.Arg1197Trp. Subsequent screening of these four variants in a case-control series consisting of 4483 individuals with idiopathic PD and 3865 age, gender, and ethnically matched control subjects identified 2 additional unrelated, affected individuals heterozygotes for the p.Ala502Val variant and 2 affected individuals from the same family both carrying the p.Gly686Cys variant. Segregation analysis was not possible in any of these cases and the two remaining variants were not observed again in affected or control subjects.

Given that the authors own data suggest an extremely low frequency for these putative causal variants within the PD population (0.2% for p.R1205H and 0.06% for p.A502V and p.G686C), assignment of pathogenicity can be difficult and the possibility that these changes are, in fact, rare but benign variants with no role in the aetiology of PD cannot be discounted. In light of this, we screened 150 familial PD cases from our UK familial Parkinson's Disease series, in which we mutations in *LRRK2*, *VPS35* and *SNCA* mutations in order to determine whether we could provide further evidence that this gene is indeed a PD-related locus [476, 477]. We also assessed these coding positions in a set of African samples (Table 38) from the Human Diversity series – a standard panel of African samples, as African samples have the greatest diversity and offer a rapid route to the identification of benign polymorphisms [478, 479].

15.2 Subjects, Materials and Methods

15.2.1 Selection of DNA Samples for Mutational Screening

The study included DNA samples obtained from 150 familial PD probands. The study was approved by the East Central London Research Ethics Committee 1 and informed consent was obtained as per its guidelines.

All living patients fulfilled criteria for clinical diagnosis of PD at the time of the study with at least 2 of 3 cardinal signs of tremor, rigidity, and bradykinesia, as well as a

positive response to levodopa therapy. Familial cases were defined as those reporting 1 or more first degree relatives with PD with a pedigree consistent with autosomal dominant inheritance and negative testing for known mutations in *SNCA* and *LRRK2*.

In addition, DNA was obtained from the Human Diversity Series, a standard panel of 127 African samples derived from cell lines. The number of samples per population and the geographic origins of the population are shown in the table below:

Table 38 – Characteristics of the DNA samples obtained for various African populations via the Human Diversity Series.

No of Samples	Population	Geographic Region
35	Biaka Pygmy	Central African Republic
15	Mbuti Pygmy	Democratic Republic of Congo
12	Bantu N.E.	Kenya
7	San	Namibia
26	Yoruba	Nigeria
24	Mandenka	Senegal
8	Bantu S.E. Pedi	South Africa

15.2.2 Mutational Screening of Exons 8 and 22 of EIF4G1

Primer pairs were designed specifically to amplify these exons of *EIF4G1* based on transcript NM_182917.3. Sequencing was accomplished as per sections 4.6 – 4.11.

15.2.3 Extraction of NHLBI Exome Sequencing Data

To obtain a more exhaustive description of the pattern of variability in EIF4G1, we also extracted genotype data from the NHLBI exome sequencing project (Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA; http://evs.gs.washington.edu/EVS/ [accessed January 2011]), which includes exome data for 3500 American individuals of European descent and 1850 African Americans. Frequencies were computed using VCFTools and annotation of variants was achieved by ANNOVAR. Both pieces of software are freely available online.

15.3 Results

15.3.1 Mutational Screening of Exons 8 and 22 of EIF4G1

Screening of exons 8 and 22 failed to reveal any mutation previously reported to have been associated with PD in our cohort of familial PD. We identified one coding change (c.1456C>T; P486S) in 2 PD patients, but this was recorded in dbSNP (rs112545306) and was observed in the NHLBI Exome Sequencing Project dataset at a frequency of 0.15% in African Americans.

In the Human Diversity Series panel of African individuals, we identified six non-synonymous changes in exons 8 and 22 of *EIF4G1*. Of these, only one was novel (c.1145C>T; P382L), whilst the rest were recorded in dbSNP and mainly seen in African populations in the NHLBI Exome Sequencing Project dataset. In order to predict the impact of on protein function of these non-synonymous variants, we performed an in-silico analysis using SIFT and PolyPhen. All varaints, including the novel variant, were predicted to be benign. These results are summarised in table 39 below:

Table 39 – Characteristics of the variants in EIF4G1 detected in the African Diversity panel by Sanger sequencing. dbSNP accession number, minor allele frequency (MAF), the population of origin and an silico prediction of pathogenicity from SIFT are provided.

Nuclotide change	Protein change	SNP accession number	MAF	Population	Predicted effect (SIFT)
c.870G>A	p.M290I	rs144947145	0.01	San, Bantu SE	Tolerated
c.913C>T	p.R305C	rs116508885	0.01	Yoruba	Tolerated
c.932A>G	p.Y311C	rs16858632	0.03	Bantu SW, Bantu NE, Yoruba, Mandenka	Tolerated
c.1145C>T	p.P382L	NA	0.004	San	Tolerated
c.1429G>A	p.E477K	rs145228718	0.004	Mandenka	Tolerated
c.3918G>A	p.R1216H	rs34086109	0.004	Biaka pygmy	Tolerated

Analysis of the NHLBI datasets demonstrated that the putatively pathogenic A502V published by Chartier-Harlin and colleagues was seen in two European-American individuals (frequency of 0.02%). In total, we identified 95 nonsynonymous SNPs spread across all 32 exons of the gene. (NM_182917.3). Of note, 36 of them are located in exon 8 or exon 22.

To further investigate coding variability across the *EIF4G1* gene we extracted the genotypes from the NHLBI dataset and computed the average number of pairwise amino acid differences between two randomly selected European-American haplotypes from the NHLBI dataset. This analysis was carried out entirely by Dr Arianna Tucci and its results are presented here only for the sake of completeness.

On average two such *EIF4G1* sequences diverge by 0.45 aminoacids: 82% of this variability (0.37 amino-acid differences) is attributable to exon 8, where the A502V lies; whilst 9.8% of this variability is attributable to exon 22 (0.098 amino-acid differences).

15.4 Discussion

Using a combination of linkage analysis and candidate gene sequencing, Chartier et al. (2011) recently reported that the p.R1205H mutation in exon 22 of the gene, *EIF4G1*, segregated with disease in a large multi-incident French family showing autosomal dominant inheritance of PD [475]. By screening additional patients with clinical or pathologically defined PD, they identified 4 less frequent, putatively disease-causing mutations, p.A502V, p.G686C, p.S1164R, and p.R1197W, which were absent from approximately 4000 controls. However, segregation analysis was not possible for these variants and thus their involvement in disease pathogenesis remains uncertain.

We screened 150 familial PD probands for the two most common mutations (p.R1205H and p.A502V) and failed to identify either of these two putative variants. Nor did we identify any other coding change that was likely to be pathogenic. Our study of the NHLBI dataset did, however, identify the p.A502V variant in 2 European-American individuals. Whilst the nature of the NHLBI dataset means that the possibility that these individuals either had or would go on to develop PD cannot be

excluded, this finding does cast some doubt on presumed pathogenicity of this variant and raise the possibility that it is, in fact, a rare benign polymorphism.

Finally, by computing pairwise amino acid differences for randomly selected European-American haplotypes from the NHLBI, Dr Tucci was able to show that there appears to be more limited selective pressure and a higher sequence variability in the regions of the *EIF4G1* protein encoded by exons 8 and 22. This finding is strengthened by the relatively high number of missense variants (n=6) identified in these exons in the African samples of the Human Diversity series.

In summary, the variants in *EIF4G1* identified by Chartier-Harlin et al. are either an extremely rare cause of PD in caucasian cohorts or, quite possibly, represent nothing more that rare benign polymorphisms in a naturally less conserved portion of the gene.

CHAPTER 16:

Summary and Discussion

16. Summary and Discussion

16.1 Preamble

In the work presented in this thesis, I sought to apply modern techniques and technology to understand the genetic basis of hereditary neurological disease. The bulk of my work focused on the use of whole exome sequencing (WES) as a tool for gene discovery in previously 'unsolved' kindreds with ostensibly Mendelian inheritance of neurological disease. In the main, the work that I have chosen to present as part of this thesis centres around kindreds exhibiting familial forms of dystonia. This is partly by design (dystonia is a condition that has always interested me and one of my primary clinical collaborators, who helped in the identification of 'unsolved' kindreds, Prof. Kailash Bhatia, is a movement disorders specialist with a particular interest in dystonia) and partly by chance (for various reasons – some of which I will touch on briefly later – it was my work on these unsolved the dystonia kindreds that turned out to be the most fruitful, whilst my work on kindreds with various other conditions tended not to advance to a stage that might be worthy of inclusion in a thesis).

This period of research has given me an in-depth understanding of the strengths and, more importantly perhaps, the challenges, pitfalls and limitations of WES in relation to gene discovery. Initially, I would like to begin by briefly summarising the successes of WES with respect to the advances expected from the advent of this technology – as previously set out in section 1.6. This will be presented both from the perspective of both my own work and from that of the field of genetic research as a whole.

16.2 WES: Promise and Reality

16.2.1 New Gene Discovery: Experience of the Field

One of the main promises of WES was that its application to previously unsolved kindreds would result in the discovery of numerous new causal genes for Mendelian disease. To a large extent, this has been true. At the time of writing, a PubMed search limited to the phrase "whole exome sequencing" alone reveals that a total of 1580 articles have been published on the technique since 2010, with the yearly totals following an roughly exponential growth pattern such that the number of publications

in 2014 alone account for nearly two thirds of this total. At the very least, this is evidence that the technique has been rapidly taken up by researchers and has now become something of a mainstay in genetic research. This rapid take up has no doubt been influenced by clear evidence of success: a recently published review article estimated that over 150 new genes causing Mendelian disease had be identified by means of next-generation sequencing [480]. Whilst this clearly a cause for general optimism, my own experience of reading the literature suggests that the distribution of gene discoveries between various forms of Mendelian disease has not been even.

The combination of homozygosity mapping and exome sequencing has proved to be particular effective in elucidating the cause of Mendelian diseases in consanguineous kindreds exhibiting recessive inheritence, even when applied to single, small families. Homozygoisty mapping tends to restrict the genomic search space to a far greater extent than is possible with linkage analysis, as generally applied to dominant or non-consanguineous recessive kindreds. Exome sequencing then functions as essentially nothing more than a quick, cheap and efficient means of sequencing the entirety of the implicated regions. To take my own work on the DYT2 kindred as an example, homozygosity mapping was able to narrow the genomic search space down from >20,000 genes to a mere 285 genes, which exome sequencing demonstrated to contain a sum total of 324 variants. After appropriate filtration, only 2 potentially causal variants remained. Given that the family consisted only of three affected siblings and their unaffected parents, the power inherent in the marriages of these two techniques is obvious. In fact, performing a PubMed search on 'whole exome sequencing' and 'homozygosity mapping' results in 114 articles, the vast majority of which purport to link a particular recessive Mendelian disease with homozygous mutations in a particular gene.

The other group of Mendelian disorders that have yielded with relative ease are those that might be best categorised as the 'very rare'. In using this label, I am thinking of disorders where one deals not so much in families, but instead in a small number of scattered individuals – as in Kabuki syndrome [84], Freeman-Sheldon syndrome [481], or static encephalopathy with neurodegeneration in adulthood [482] (now renamed beta

propeller protein associated neurodegeneration, or BPAN for short, in light of the discovery of the causal gene). Many of these 'very rare' disorders arise from *de novo* mutations and linkage analysis is thus less useful, particularly if there is unsuspected genotypic heterogeneity. However, by exome sequencing trios (both unaffected parents and the affected child), the number of potentially causal variants can be rapidly reduced. Due to the intrinsic error rate of DNA polymerase, each individual carries about 70 de novo variants with respect to their parents (see section 1.8). Most of these can be rapidly excluded on the basis that they do not affect protein coding. It is then a relatively simple step to identify potentially causal, *de novo* variants that are seen in many or most affected individuals or, alternatively, to identify genes that are affected by multiple different, potentially causal variants in the case cohort as a whole (see section 1.5.2 for some caveats to this statement). Phenotypic stratification can (and, indeed, has been used to) help guide researchers in identifying such genes where genotypic heterogeneity initially obscures the results. Indeed, this was the approach used to identify the first Mendelian disease-gene for Kabuki syndrome in 2010 [84]. As a rough guide to success, a PubMed search for 'whole exome sequencing' and 'de novo mutation' results in 153 articles, of which approximately half purport to explain various very rare Mendelian diseases in terms of *de novo* mutations in some gene.

The disorders where WES has proved least fruitful are the dominantly inherited Mendelian disorders, particularly those with a late onset or where reduced penetrance or phenotypic heterogeneity is a factor. The unexpectedly high level of genetic variation encountered in the average healthy exome, including numerous potentially and frankly deleterious changes (see section 1.4), means that the exome data alone is rarely ever sufficient to pinpoint the likely causal variant, even when multiple affected and unaffected individuals in a kindred are sequenced. Some means of narrowing the genomic search space is always required and, for dominant disorders, this usually means linkage analysis. Whilst, as my work on *ANO3* shows, the use of exome sequencing does mean that there is no longer the requirement to identify a single tight region of linkage with a truly significant LOD score, analysis must still result in an area or areas of positive linkage that cover a sufficiently small number of potentially causal variants to make further progression to segregation analysis and sequencing in

independent cohorts a feasible option. Most of the largest dominant kindreds for Mendelian disease have already been 'solved', either through positional cloning or by virtue of testing positive for a gene already discovered by that means in some other family. Given the tendency to small family size in the Western world, most of the remaining unsolved kindreds are small or, at best, moderately sized, usually with a total of less than 10 individuals in the three living generations (if one assumes ~2 children per generation, on average). With respect to the disease under study, a later age at onset, reduced penetrance and/or phenotypic heterogeneity can have a large impact on linkage analysis. Later age at onset often means that the youngest generation(s) cannot be included in any analysis since it is impossible to say whether they are truly unaffected or whether they will merely manifest at a later date. Reduced penetrance can dramatically reduce the power of linkage analysis in small families, when a non-manifesting carrier is included as an unaffected individual, leading to spurious areas of positive linkage that swell the list of potentially causal variants to unmanageable levels. Phenotypic variation has a similar effect, but can also lead to the labelling of affected individuals as unaffected if their phenotype is so mild as to go unnoticed/unreported or sufficiently different to be thought to represent an unrelated medical condition.

16.2.2 My Own Experience of New Gene Discovery: The Good

Within this thesis, I have presented work that demonstrates a range of outcomes resulting from the application of WES to unsolved Mendelian kindreds. This work has resulted in the publication of two new genes for primary isolated dystonia, *ANO3* (DYT24) and *HPCA* (DYT2). In addition, the work presented in chapter 9, makes a compelling case – to my mind at least – that the homozygous mutations that we detected in the gene *SLC25A46* are the most likely cause of the complex neurological phenotype detailed therein, even if I do not currently have sufficient evidence for publication.

ANO3

ANO3 was identified in a moderately-sized Caucasian kindred exhibiting autosomal dominant inheritance of tremulous craniocervical dystonia. There is little doubt that the structure of the family brought with it some advantages from the point of view of a

traditional linkage analysis approach to gene identification: the index affected individual had remarried and transmission of the causal mutation to subsequent generations had occurred on both sides of the family resulting in a doubling of the family size by Western standards (and, by extension, in the availability of affected individuals with greater genetic separation for subsequent exome sequencing). In the end, however, even this fortuitous pattern of inheritance was not sufficient to produce a truly significant linkage peak. Instead, analysis yielded 5 linkage peaks with a LOD score of 2.01. In total, these linkage peaks covered a total of 801 known genes. Positional cloning would thus clearly have been a non-starter. Even a candidate gene approach would not have helped. Based on homology to *ANO1* and *ANO2*, the best guess at the time of the work on this family was that *ANO3* might function as a Ca2+-activated chloride channel. A gene thought to function as an ion channel is unlikely to have been prioritised given the state of knowledge at that time of the pathogenesis of dystonia at that time. (Since then, the assumption that *ANO3* functions as an ion channel has actually been challenged [483] and the best that can be currently be said of its true function is simply that it remains unknown at present).

In total, using a targeted NGS approach, we identified five potentially causal mutations in this gene in separate individuals with craniocervical dystonia. In two of these individuals, autosomal dominant inheritance was evident within their living family members and, in both – admittedly small – kindreds, the mutations in *ANO3* showed perfect segregation with the disease phenotype. The phenotype of disease in these families matched that of the index family perfectly: tremulous, slightly jerky craniocervical dystonia with its onset in the first five decades of life and with no generalisation despite prolonged follow-up.

Like all new genes, the true test of the validity of *ANO3* as a new dystonia gene will be the identification of further kindreds exhibiting dystonia that segregates with mutations in the gene can by independent research groups. In this regard, *ANO3* has one particular disadvantage: its size. *ANO3* is composed of 27 coding exons making mutational screening by Sanger Sequencing in any significant number of cases a very time consuming and expensive process. It is likely that screening a gene of this size on

any large scale could only be untaken by NGS, most likely by using a targeted approach aimed at simultaneously sequencing all reported dystonia genes. Whilst this approach is certainly feasible today and has already been used for mutational screening in other disorders, it is yet to be applied in the field of dystonia research, at least in any publication that I currently aware of. The wait for conformation may therefore be protracted.

HPCA:

HPCA was identified in a small consanguineous kindred of Sephardic Jewish ancestry who exhibited young-onset generalised dystonia with a craniocervical predominance, inherited in an ostensibly recessive manner. In order to identify the gene, a combination of homozygoisty mapping and exome sequencing were applied to the family. As stated above, this combination represents a powerful technique in genetic discovery since homozygosity mapping requires far fewer individuals DNA to effectively narrow the genomic search space and can usually do so to a far greater degree than does linkage analysis, unless the kindred is very large. In the case of this particular family, after appropriate filtration, only 2 candidate causal variants were identified within the largest tract of homozygosity for further consideration. In trying to determine which of these two variants was most likely to be causal, we were helped considerably by the extremely favourable expression pattern of *HPCA* compared with that of *LAPTM5* and also by the fact that the homozygous mutation in the index family affected a Ca^{2+}-coordinating amino acid in the critical EF hand domain 2, causing the substitution a positively-charged lysine residue in place of the widetype neutral asparagine. Binding of a positively-charged Ca^{2+} to this domain is required to operate the mirystoyl switch mechanism that results in a conformational change in the protein that causes extrusion of the mirystoyl moiety. The extrusion of this moiety allows the protein to translocate to the plasma membrane and bind with downstream interactors. The nature and position of the mutation that we detected thus suggested a plausible mechanism for pathogenicity. The finding of a further kindred with compound heterozygous mutations in this gene was perhaps lucky given the apparent rarity of autosomal recessive isolated dystonia and the relatively low level of natural sequence variation in publically available databases. Despite this, one might reasonably hope that

identification of further independent kindreds by other research groups will be considerably easier than for *ANO3*. The gene is only very small – consisting of just 3 coding exons – and will thus be easy to screen in large cohorts using even conventional methodologies. Furthermore, whilst the autosomal recessive isolated dystonia probably is very rare in populations where consanguineous marriages are infrequent, such as that of most of Europe and the USA, it probably exists at a significantly higher frequency elsewhere and may thus turn out to be more common globally than expected.

Finally, in chapter 9, I presented my work on a small consanguineous kindred exhibiting recessive inheritance of a complex neurological phenotype consisting of bilateral visual failure, severe action myoclonus, and a peripheral motorsensory axonal polyneuropathy. Bilateral visual failure, particularly in the context of clinical evidence to suggest widespread peripheral and central neurological dysfunction, is highly suggestive of mitochondrial dysfunction. With this in mind, the novel, homozygous mutation in the gene *SLC25A46*, which encodes a mitochondrial carrier protein highly expressed in the nervous system, naturally presented itself as a top candidate on the basis of its inherent biological plausibility. Moreover, the affected amino acid forms part of the key solute carrier domain of the gene's protein product. Unsurprisingly perhaps, it is thus highly conserved and its substitution was universally predicted to be deleterious by all *in silico* prediction methods used. Indeed, it is my belief that this is the causative mutation in this family. As of yet, however, I do not yet have enough evidence to publish. Ideally, I would have liked to find a second independent individual or kindred with a similar phenotype who possessed biallelic mutations in this gene. Unfortunately, the phenotype is quite unusual and we were unable to identify any cases that matched the clinical description at our institution.

Recently, two possible means of advancing with work on this kindred have come to light. Contrary to the information that I was first given, the family now recall that there is, in fact, a cousin who also had a child who died in infancy of a severe neurological disease. I am currently trying to obtain a clinical description of the child's symptoms and signs, but it is not unreasonable to hypothesise that their illness may represent a more severe presentation of the same genetic lesion. We have established

that DNA from this child is held in storage at Great Ormond Street hospital and have been provide with consent from the family to obtain a sample for genetic testing. If the homozygous mutation in *SLC25A46* can be demonstrated in the DNA of the child and, with luck, the other remaining candidates shown to be absent, this would probably be sufficient evidence to go ahead and publish on the basis of a single family. (This scenario is feasible because, apart from *SLC25A46*, no other candidate variant is located in the same stretch of homozygosity on chromosome 5. In particular, if the decreased child did not share the share the tract of homozygosity on chromosome 11, this would immediately eliminate 3 of the 4 other candidate variants)

From the point of view of assembling a phenotypically similar cohort, I have recently learned that the Neuromuscular department may keep a list of the samples from patient with a mitochrondrial phenotype that includes visual failure secondary to rod-cone dystrophy in whom no mutation has been identified. I am current attempting to make contact with the relevant parties to obtain DNA from these individuals for testing.

16.2.3 My Own Experience of New Gene Discovery: The Bad

Within this category, I include my work on kindreds that progressed to an advanced phase but failed to reveal any variant for which a reasonable argument for pathogenicity could be maintained. The two examples of such kindreds that I have included in my thesis are presented in chapters 6 and 11.

In chapter 6, I presented my work on a large kindred exhibiting ostensibly autosomal dominant inheritance of late-onset tremulous cervical dystonia. Initially, producing a workable list of candidate variants in this family was hampered by the fact that the late-onset of the disorder meant that the affectation status of only a single generation could accurately be determined. This precluded any meaningful linkage analysis and thus afforded no opportunity to narrow the genomic search space. Concessions thus had to be made from the very outset, including, most notably, filtering the exome data to leave only novel variants shared between all three affected members for further consideration. This was followed by a decision about which of the candidate variants to test for segregation in the family based on a compiled character profile. Although I

tried to be as unbiased and as scientific in going about this as possible, the process necessarily entailed some degree subjectivity. In the end, none of the variants chosen for segregation analysis could be shown to fully segregate with disease in the family, bringing work to a grinding halt. Nonetheless, it is notable that the variant that I had chosen as my top candidate – the p.Y605C variant in *CACNA2D3* – came closest of all to fully segregating with disease, raising the possibility that the single mismatch observed was due to a phenocopy within the family. Indeed, the possibility of phenocopies had been suggested to me at the outset by my supervisor as he felt that the family appeared slightly 'over-dominant'. As stated in the conclusion of chapter 6, the passage of time will probably be the most helpful factor in solving this family as accurate knowledge of the affectation status of the second generation will permit linkage analysis and increase the power of segregation analysis. Unfortunately, given the average age at onset of symptoms in affected individuals was in the 7th decade, another 20 or 30 years will probably required to pass before this becomes a reality!

In chapter 11, I presented my work on a small, consanguineous Pakistani kindred exhibiting infantile onset, autosomal recessive choreodystonia. Four potentially causal candidate homozygous variants remained in this family after appropriate filtration, of which that detected in the gene *TRIM3* seemed the most plausible based on its postion with in a key functional domain of the protein and what is currently known about the genes function and expression pattern. Unfortunately, the nature of the variant itself was far less convincing – consisting of a glycine-to-alanine substitution that was universally predicted to be tolerated – and no further potentially causal mutations could be found in a small independent cohort of young onset dystonia (though, admittedly, the phenotype of individuals included in this cohort was not a particularly good match for that of the affected individuals in the index family). It is difficult to see how work could be advanced further on this family beyond further mutational screening of all candidate causal genes in a phenotypically well matched cohort of individuals from independent kindreds, which is not currently possible at our institution.

16.2.4 My Own Experience of New Gene Discovery: The Just Plain Ugly

The kindreds presented herein are only a selection of those on which I worked during my time at the Institute of Neurology. They represent those in which the work advanced to a stage where the list of variants could be sufficiently narrowed so as not allow some vaguely scientific means of progressing towards a list of one or more top candidates. There were, however, several further kindreds where work stopped short of this point. By and large, the problem was either one of too many or, alternatively, too few variants.

In some kindreds, difficulties in accurate phenotyping (dystonia), small family size that precluded any form of linkage (dystonia, ataxia), or lack of any similar cases for screening at our institute (familial temporal lobe epilepsy) meant that the list of variants was so unwieldy that my attempts to identify top candidates was in reality nothing more than hopeful guesswork. It is possible that, with time or the development and application of other ancillary clinical techniques, some of these families will yield: small families, if followed up, will eventually become large enough to permit some form of linkage, whilst re-phenotying of some kindreds using surrogate markers of carrier status, such as temporal discrimination testing in dystonia, may prove a workable strategy in the future.

In some kindreds, on the other hand, the problem was one of two few variants. By way of example, I worked on two families where the mode of inheritance (x-linked parkinsonism with psychiatric features) or a previously-published, seemingly-robust linkage interval (geniospasm) had tempted me into thinking that finding the candidate variant might a relatively easy affair. It turned out not to be the case: no shared potentially pathogenic variants could be found in the anticipated regions. There are a number of possible explanations for this. Firstly, the limitations placed on the genomic search space could have been incorrect. The clinical manifestations of geniospasm are relatively subtle and there is a tendency for them to remit as the patient ages so that, in some cases, family members' reports that they were affected at birth were used to assign affectation status, despite the fact this could no longer by verified by clinical examination. There was thus a very real risk of incorrectly assigning affected status,

leading to incorrect linkage and a search that is doomed to fail from the beginning. In the case of the kindred exhibiting hereditary parkinsonism with psychiatric features, the inheritance pattern that best fitted the pattern of disease in the family was x-linked recessive (with no affected females and no male with affected children) but the family was relatively small and it is possible (albeit less likely) that disease could have been inherited in a autosomal dominant pattern with reduced penetrance. Whilst affectation status in this family could certainly much more easily be assigned with relative certainty, it remained too small for any useful genome-wide linkage analysis and so that, if autosomal dominant inheritance with reduced penetrance was assumed, the list of potentially causal variants once again became unmanageably large. On the assumption that the mode of inheritance is truly x-linked and the causal variant on X has merely been missed by exome sequencing, DNA from this family has been sent for whole genome sequencing and search for it has now been taken up by one of my research colleagues, Dr Niccolo Mencacci. I await the results of his analysis with interest.

16.3 Expanding the Phenotype: My Own Experience

In section 1.6.2, I stated that one of the spin-offs of widespread use of exome sequencing had been increased awareness of genetic pleiotropy and gave a few examples genes in which mutations are now recognised to cause a much wider phenotypic spectrum of disease than previously thought. For my own part, the research presented herein has led to the identification of two new phenotypes for known Mendelian disease genes, one already published and one in the process of publication.

Firstly, in chapter 8, I was able to demonstrate that compound heterozygous mutations in *ATM* were responsible for a variant form of ataxia telangiectasia (AT) in a small muslim Indian kindred. The phenotype that the affected individuals exhibited was characterised by dopa-responsive cervical dystonia and ocular telangiectasia without any associated ataxia. Although variant presentations of AT are well documented, dopa-responsive dystonia had never previously been reported as a possible manifestation of mutations in this gene, meaning AT had never been considered as a possible differential diagnosis in this family. Indeed, the mild ocular telangiectasia was only

noted once the genetic diagnosis became apparent. This discovery adds AT to the list of diagnostic considerations in undiagnosed dopa-responsive dystonia (particularly if cervical) and this is important since, despite the lack of typical features, individuals with variant presentations of AT still appear to be at the same elevated risk of malignancy and medical investigations involving radiation must be avoided.

Secondly, in chapter 10, I have presented compelling evidence to suggest biallelic mutations in the gene *NUBPL* – previously only associated with a very severe disease characterised radiologically by widespread leucodystrophy – can also result in a much milder phenotype, characterised instead by young-onset dystonia and bilateral striatal necrosis. As part of my argument, I suggest that by and large the previously-published, ostensibly 'characteristic' phenotype of *NUBPL*-related disease has resulted from systematic ascertainment bias in the selection of cases for mutational screening. This work therefore highlights the danger inherent in such an approach in an age when phenotypic pleiotropy is being increasingly recognised to play an important role in genetic disease. Although this presents a challenge to clinicians who must decide which genes to test based on the phenotype of the patient, it is to be hoped that a shift from reliance on traditional sequencing methodologies to next-generation targeted or even whole exome sequencing in genetic diagnosis will compensate for this increased complexity.

16.4 Brief Summary of Other Aspects of My Research

In addition my use of exome sequencing to attempt to discover new Mendelian disease genes, I also used more traditional sequencing methods to confirm the validity and assess the prevalence of Mendelian disease genes published by others, namely *VPS35*, *GNAL* and *EIF4G1*. In addition, in a brief break away from the world of Mendelian genetics, I performed a small association study in neuropathologically proven PD. These endeavours are briefly summarised in the following sections.

16.4.1 VPS35 is a Rare Cause of Familial Parkinson's Disease

As with all genes identified by NGS, including those I have proposed herein, it is important that replication studies are performed in order to confirm the result in other populations; to determine whether the mutations reported in the index study are truly pathogenic; to assess the prevalence of mutations in that gene as a cause of disease; and to assess whether there are any phenotypic clues to the underlying genetic cause that may be of use to clinicians.

Mutations is the gene *VPS35* were reported to cause autosomal dominant Parkinson's disease (PD) by two groups simultaneously in 2011 [466, 484]. In collaboration with a fellow researcher, I subsequently performed mutational screening of the gene in a large UK Parkinsons Disease. We identified one large kindred split between the UK and Australia that harboured the already published p.Asp620Asn mutation, further confirming the validity of VPS35 as a Mendelian disease gene for Parkinson's disease. The phenotype of the kindred was essentially indistinguishable from that of late-onset idiopathic Parkinson's disease

16.4.2 GNAL is not a Common Cause of Primary Isolated Dystonia

Mutations is the gene *GNAL* were reported to cause autosomal dominant primary isolated dystonia in early 2013 [334] and confirmed by a second group shortly thereafter . The phenotype was said to consist of cervical predominant dystonia with generalisation in about 10%. The initial discovery paper suggested a somewhat incredible prevalence of approximately 15%, with the authors claiming to have detected mutations in 6 of only 39 families screened. If verified, this claim would have significant implications for genetic testing, since cervical dystonia is common and testing for *GNAL* mutations would have seemed mandatory in anybody with a family history or a younger age at onset. I set about screening 196 individuals with familial and sporadic cervical onset or predominant dystonia. I detected no pathogenic mutations in this cohort and argued that *GNAL* was very unlikely to be a common cause of cervical dystonia, in the UK population at least [485]. Subsequent screening studies in other populations have

supported this assertion, suggesting the true prevalence is likely to be closer to 0.5 – 1% [486].

16.4.3 The Link Between EIF4G1 and PD is Tenuous

Mutations in EIF4G1 were published as a cause of autosomal dominant PD in September 2011 [475]. In collaboration with two other research colleagues at our institution, I performed mutational screening of the gene in a cohort 150 UK familial PD cases. We did not find any pathogenic mutations in this cohort. In addition, we demonstrated that the one of the proposed pathogenic variants from the original study (p.Ala502Val) was in fact present in the datasets of the NHLBI Exome Sequencing Project, casting significant doubt on its pathogenicity. Subsequent studies of a much larger scale have confirmed that *EIF4G1* is not a high risk PD gene and that the originally published p.Arg1205His mutation is not significantly associated with increased risk of PD, in the Icelandic population at least [487].

16.4.4 MAPT is Associated with Neuropathologically-Proven Parkinson's Disease

Association studies in Parkinson's disease have demonstrated multiple genetic loci that influence risk in the idiopathic form of the disorder. Prior to my work, however, all previous studies had been carried out using samples from clinically diagnosed cases. As I point out in chapter 12, however, the error rate for clinical diagnosis of idiopathic Parkinson's disease is in the range of 10 – 25% [452]. Given that at least one of the common clinical mimics, progressive supranuclear palsy (PSP), is not a synucleinopathy at all but a tauopathy instead and that the relevance of tau aggregation in Parkinson's disease was uncertain, we set about answering the question of whether the previously reported association in *MAPT* might be due to contamination of the case group by individual's with PSP who had been misdiagnosed as having idiopathic Parkinson's disease My work showed that the association signal in *MAPT* remained even when the case cohort was defined purely by neuropathological means and simultaneously confirmed the association with the well known *SNCA* locus. This suggests tau aggregation has a role to play in the pathogenesis of idiopathic PD.

16.5 Future Directions in Next Generation Sequencing

By way of conclusion, I would like to briefly consider the developments that are likely in genetic research into the causes of human disease in the near future, as familiarity with NGS techniques increases and cost continues to fall. The rewards in terms of the advancement of scientific understanding promise to be significant. At the same time, however, as I highlight in the following sections, trail-blazing pioneers are likely to face considerable challenges along the way.

16.5.1 The Argument for Whole Genome Sequencing

The widespread application of whole exome sequencing to the genetics of Mendelian disease has resulted in a rapid acceleration of the pace of discovery. More than 85% of known disease-causing mutations occur in the protein coding regions of the genome (i.e. exons) [20]. Sequencing of this part of the genome thus provides knowledge about genetic variation in a region that should be highly enriched for variants with consequence for human health, making the identification of clinically meaningful results more efficient. However, the exome represents only about 1% of the entire genome and it should be recognised that to say that 85% of known disease-causing mutations occur in protein-coding regions of the genome is not equivalent to saying that the 85% of disease-causing mutations will be found to affect protein-coding DNA sequences. I think it would be incredibly short-sighted to believe that the remaining 99% of the human genome will not, in the future, be shown to be important in the causality of more than the currently estimated 15% of human disease. (It should be noted also that much of the current estimate of the 15% contribution of non-coding regions to Mendelian disease is taken up my non-coding variants affecting canonical splice sites immediately adjacent to exon/intron boundaries, which can, in any case, be captured by current exome sequencing techniques). Perhaps the most notable recent example of the identification of non-coding genetic variation as a cause of Mendelian disease is encapsulated by the discovery of a intronic expansion in the gene C9orf72 as a relatively common cause of both familial amyotrophic lateral sclerosis and frontotemporal dementia. This discovery signposts the possibility that that non-coding variation may turn out to have a much more important role – numerically speaking – in human disease than has been anticipated to date. In fact, it is currently thought that

the the C9orf72 mutation is the single most common genetic cause of ALS and FTD, accounting for up to 50% of familial cases in and over 25% of sporadic cases in some European populations [80, 81].

Whole genome sequencing is technically feasible today and would thus seem to be the logical successor to whole exome sequencing in research aimed at discovery of new causes of Mendalian disease. Certainly, it would offer a number of theoretical advantages over whole exome sequencing. Most obviously, it would allow for a complete assessment of the genetic variation in relation to Mendalian disease. In theory at least, disease resulting from non-coding variants could be identified by demonstrating segregation in much the same manner as is currently done in experiments utilising whole exome data. Moreover, whole genome sequencing would allow for the assessment of the contribution of exonic copy number variants and large-scale deletions in a manner that is currently only truly possible by the use of ancillary genetic techniques, such as multiplex ligation dependent probe amplification or fluorescent in-situ probe hybridisation, respectively.

16.5.2 Super Size Me: Giant Haystacks of Whole Genome Data

Despite the possible benefits, the idea that whole-genome sequencing data could really be used in the same manner that whole-exome data is used today is currently somewhat far-fetched. Some have likened identifying the causal variant amidst the background of natural human variation in an exome to finding a needle in a haystack. However, the average human exome typically includes a mere ~22,000 variants; in comparison, the average human genome contains around 4 million variants, of which about 3.5 million are single nucleotide variants [488]. The haystack just got a whole lot bigger! Two major problems arise from working with the entire human genome: the first relates to a relative lack of annotation; the second to the difficulty in prioritising candidate variants.

16.5.3 Whole Genome Data: The Pressing Need for a Better Map

The challenge of interpretation of non-coding sequencing data is made significantly harder by an imperfect knowledge of the full range of mechanisms of regulatory control

and the incomplete annotation of even those that are recognised. Currently recognised regulatory mechanisms that have been associated with human disease are diverse and include: splice sites (both canonical and weakly specified) [489-491]; sequences regulating translation, stability and localisation (5' and 3'UTRs) [492, 493]; trans-regulatory RNAs (i.e. long non-coding RNA) [494, 495]; promotor sequences [496-498]; enhancer sequences [499]; and synonymous mutations within coding sequences [500]. Several large-scale projects are currently underway to enhance the annotation of the non-coding genome, which will be a necessary first step in permitting systematic interpretation of variants in these regions.

16.5.4 Whole Genome Data: The Problem of Variant Filtration

To date, almost all successful disease gene discovery studies using exome sequencing have relied on a form of variant filtration designed to whittle down the relatively modest number of variants detected by such studies to a small list of one or more potentially causal variants. This filtration process relies on knowledge of which variants are observed in the normal healthy population, which positions have been conserved through evolution by purifying selection and, occasionally, on the ability of tried and tested *in silico* prediction programs to give an indication of the likelihood that a variant will have a functional consequences on human physiology. Unfortunately, at present, this knowledge base simply does not exist for the entire human genome.

Firstly, the number of human genomes to be sequenced in their entirety for which data has been made publically available remains relatively modest. The 1000 Genomes Project probably represents the greatest resource at present. By definition, however, this can only hope to document human genetic variation down to a MAF of 0.1%, which, in reality, only encompasses relatively common variations (the definition of a polymorphism is a variant occurring at a frequency of 0.5% or greater). Moreover, the number number of populations currently sampled is limited. Given that a significant proportion of variants detected by exome sequencing are rare or, even, effectively private to particular sub-population or even individuals, there is no reason to suspect that a similar situation will not prevail in the case of whole-genome data. In fact, there is good reason to believe that a greater toleration of genetic drift in non-coding regions

would mean that the extent of rare or semi-private variation in small sub-populations would be far greater in non-coding regions than is seen in protein-coding regions.

Secondly, the use of evolutionary conservation as a guide to the likely deleteriousness of an observed variant relies on genomic sequencing data being made available for a significant number of other species. Naturally, genomic sequencing in other species tends to lag behind that of humans (with the exception of well-studied creatures, such as the worm and fly, that possess smaller, more easily sequenced genomes). Moreover, natural variation between species is far greater in non-coding regions than in coding regions, making identification of orthologous DNA sequences and multispecies alignment much harder. Taken together, these facts make estimation of the true degree of evolutionary conservation more difficult and more prone to error on a genomic level.

Finally, since both the degree of evolutionary conservation at the site of a variant and its appearance in databases of normal sequence variation, in combination with the physicochemical properties of the observed amino acid substitutions, are some of the main considerations taken in account by algoritms making *in silico* predictions of pathogenicity, most of the currently available prediction programs are currently unable to provide any predictions for non-coding substitutions. Even in the case of those that can (such as MutationTaster), the accuracy of their predictions is far from certain. As a consequence, there is currently no real tried and tested method of predicting the functional consequence of any given non-coding variant on human physiology. Whilst the imperfections of the predictions SIFT and PolyPhen are well documented (and are briefly summarised in sections 2.7.4 – 2.7.7), they do, at the very least, provide a helpful guide as to the likely functional consequence of a given protein-coding variant, which will not be available to researchers picking over non-coding data. Analytical restriction of *in silico* predictions of pathogenicity to coding mutations would seem, therefore, to be untenable in the long term.

One final practical consideration that needs to be taken into account is the sheer volumes of data generated by whole genome sequencing. This will produce significant

challenges in terms of data storage and bioinformatics analysis for laboratories working with whole genome data on any sort of scale. For individual researchers, it is notable that the storage space required for a raw sequence data file from a single whole genome currently exceeds the capacity of most desktop computers. This means that systems will have to be developed that allow remote, seamless access to sequencing data stored on enormous central servers.

16.5.3 Association Studies Using Rare Variants

Although genome wide association studies (GWAS) have successfully revealed numerous susceptibility genes for common neurologic diseases, the odds ratios associated with these risk alleles are relatively low and, even taken together, account for only a small part of the estimated heritability of the condition being examined. This state of affairs has led to the hypothesis that a significant proportion of this missing heritability may be accounted for by rare variants of moderate to large effect size, affecting both coding and non-coding regulatory sequences. Chip-based GWAS, however, can only detect common genetic variation as rare variants are poorly targeted by design. The advent of next-generation sequencing, on the other hand, and the rapid precipitous fall in its cost have paved the way for the detection of rare, disease-associated variants via exome or even whole genome sequencing, in theory at least. Although association studies based on whole genome data would allow an assessment of the contribution of nearly all forms of genetic variation to human disease, it is unlikely that any such studies will be attempted any time soon. The cost remains prohibitive when one considers the number of individuals that must be sequenced to perform an adequately powered association study and, as mentioned earlier, handling, storage and interpretation of such vast quantities of data would be problematic. Whilst association studies based on exome sequencing data can only ever allow an assessment of the contribution of rare coding variants to disease risk, they are far more attractive from a financial and technical view point and I suspect it will not be long before they begin to appear in the publication record. That said, even for the use of exome data in this manner, there remain considerable challenges that will have to be tackled before such studies might stand a chance of success.

One of the most significant challenges will be obtaining adequate power to detect the effects of such rare variants. In general, the power of association tests between single variants and a disease trait decreases as the minor allele frequency of the variant does. A large effect size can compensate for this to some degree, but only if it is very large. Thus, rare variant association studies are likely to require enormous sample sizes, especially it there is to be any hope of detecting those rare variants with only moderate effect sizes. This will significantly increase the cost of such studies and will mandate large, international cooperative efforts to achieve the necessary sample sizes.

A second equally challenging problem is that of defining the level of association that will be considered significant. The typical chip-based GWAS significance threshold is set at 5×10^{-8}, the result of a statistical correction that is based on the assumption that approximately one million independent tests are being conducted simultaneously. Whilst significant linkage would mean that tests conducted on rare variants are not really completely independent, it is, nonetheless, likely that there will be far more rare variants than there are common variants genotyped by DNA array in present-day GWAS. This would mean a required threshold of significance for rare variant association studies that is even more stringent than that currently employed in chip-based GWAS. Such a stringent threshold may be unattainable in a single variant analysis, unless the sample sizes are truly enormous. One solution to this problem may be to analyse multiple rare variants within a given region together, with the most convenient division being a gene. The most advanced application of these methods to date can take into account both risk and protective variants within the same region, aggregating the individual association signals into a single combined association score [76, 501].

A final problem arises from the fact that many rare variants are of recent origin in human evolutionary terms and, as a result of a rapid human demographic explosion, are often confined to particular subpopulations (see section 1.4) [502]. This makes rare variant studies more susceptible to inter-population differences in allele frequency and false positive associations can more easily result from subtle underlying population

substructure [503]. Whilst robust methods have been developed to correct for population stratification in GWAS, they are currently lacking for rare variant association studies.

In summary, although the use of NGS data in association studies is now feasible and intrinsically attractive, there remain considerable challenges to its implementation and many prerequisite methodologies remain in their infancy.

17. References

1. Suzuki Y. Molecular basis of neurogenetic diseases. Brain Dev 2011;33:719-725.
2. Gerrish A, Russo G, Richards A, et al. The role of variation at AbetaPP, PSEN1, PSEN2, and MAPT in late onset Alzheimer's disease. J Alzheimers Dis 2012;28:377-387.
3. International Parkinson's Disease Genomics Consortium, Nalls MA, Plagnol V, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet 2011;377:641-649.
4. International Parkinson's Disease Genomics Consortium, Wellcome Trust Case Control Consortium 2. A Two-Stage Meta-Analysis Identifies Several New Loci for Parkinson's Disease. PLoS Genetics 2011;(Currently in publication).
5. Pankratz N, Wilk JB, Latourelle JC, et al. Genomewide association study for susceptibility genes contributing to familial Parkinson disease. Hum Genet 2009;124:593-605.
6. Kaiser J. Human genetics. Affordable 'exomes' fill gaps in a catalog of rare diseases. Science 2010;330:903.
7. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94:441-448.
8. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics 2003;33 Suppl:228-237.
9. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 2009;106:19096-19101.
10. Coventry A, Bull-Otterson LM, Liu X, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 2010;1:131.
11. Li Y, Vinckenbosch N, Tian G, et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nature Genetics 2010;42:969-972.
12. Marth GT, Yu F, Indap AR, et al. The functional spectrum of low-frequency coding variation. Genome biology 2011;12:R84.
13. Nelson MR, Wegmann D, Ehm MG, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 2012;337:100-104.
14. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64-69.

15. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 2007;80:727-739.
16. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics 1999;22:231-238.
17. Zhu Q, Ge D, Maia JM, et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American journal of human genetics 2011;88:458-468.
18. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature 2011;475:163-165.
19. Gravel S, Henn BM, Gutenkunst RN, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 2011;108:11983-11988.
20. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011;12:745-755.
21. Xue Y, Chen Y, Ayub Q, et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. American journal of human genetics 2012;91:1022-1032.
22. Wider C, Melquist S, Hauf M, et al. Study of a Swiss dopa-responsive dystonia family with a deletion in GCH1: redefining DYT14 as DYT5. Neurology 2008;70:1377-1383.
23. Weber YG, Kamm C, Suls A, et al. Paroxysmal choreoathetosis/spasticity (DYT9) is caused by a GLUT1 defect. Neurology 2011;77:959-964.
24. Kamphans T, Sabri P, Zhu N, et al. Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PLoS ONE 2013;8:e70151.
25. Petukhova L, Shimomura Y, Wajid M, Gorroochurn P, Hodge SE, Christiano AM. The effect of inbreeding on the distribution of compound heterozygotes: a lesson from Lipase H mutations in autosomal recessive woolly hair/hypotrichosis. Hum Hered 2009;68:117-130.
26. Benayoun L, Spiegel R, Auslender N, et al. Genetic heterogeneity in two consanguineous families segregating early onset retinal degeneration: the pitfalls of homozygosity mapping. Am J Med Genet A 2009;149A:650-656.
27. Lezirovitz K, Pardono E, de Mello Auricchio MT, et al. Unexpected genetic heterogeneity in a large consanguineous Brazilian pedigree presenting deafness. European journal of human genetics : EJHG 2008;16:89-96.
28. Laurier V, Stoetzel C, Muller J, et al. Pitfalls of homozygosity mapping: an extended consanguineous Bardet-Biedl syndrome family with two mutant genes (BBS2, BBS10), three mutations, but no triallelism. European journal of human genetics : EJHG 2006;14:1195-1203.

29. Miano MG, Jacobson SG, Carothers A, et al. Pitfalls in homozygosity mapping. American journal of human genetics 2000;67:1348-1351.
30. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics 2000;156:297-304.
31. Vissers LE, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. Nature Genetics 2010;42:1109-1112.
32. Girard SL, Gauthier J, Noreau A, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nature Genetics 2011;43:860-863.
33. O'Roak BJ, Deriziotis P, Lee C, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature Genetics 2011;43:585-589.
34. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 2012;335:823-828.
35. Li MX, Kwan JS, Bao SY, et al. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genetics 2013;9:e1003143.
36. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 2011;88:440-449.
37. Tchernitchko D, Goossens M, Wajcman H. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clinical chemistry 2004;50:1974-1978.
38. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers 2010;14:533-537.
39. Dorfman R, Nalpathamkalam T, Taylor C, et al. Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene? Clin Genet 2010;77:464-473.
40. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet 2012;57:621-632.
41. Stearns FW. One hundred years of pleiotropy: a retrospective. Genetics 2010;186:767-773.
42. Ishii A, Saito Y, Mitsui J, et al. Identification of ATP1A3 Mutations by Exome Sequencing as the Cause of Alternating Hemiplegia of Childhood in Japanese Patients. PLoS ONE 2013;8:e56120.
43. Aminkeng F. Mutations in ATP1A3 cause alternating hemiplegia of childhood. Clinical Genetics 2013;83:32-33.

44. Rosewich H, Thiele H, Ohlenbusch A, et al. Heterozygous de-novo mutations in ATP1A3 in patients with alternating hemiplegia of childhood: a whole-exome sequencing gene-identification study. Lancet neurology 2012;11:764-773.
45. Bras J, Verloes A, Schneider SA, Mole SE, Guerreiro RJ. Mutation of the parkinsonism gene ATP13A2 causes neuronal ceroid-lipofuscinosis. Human Molecular Genetics 2012;21:2646-2650.
46. Wang JL, Cao L, Li XH, et al. Identification of PRRT2 as the causative gene of paroxysmal kinesigenic dyskinesias. Brain : a journal of neurology 2011;134:3493-3501.
47. Chen WJ, Lin Y, Xiong ZQ, et al. Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. Nature Genetics 2011;43:1252-1255.
48. Lee HY, Huang Y, Bruneau N, et al. Mutations in the novel protein PRRT2 cause paroxysmal kinesigenic dyskinesia with infantile convulsions. Cell Rep 2012;1:2-12.
49. Heron SE, Grinton BE, Kivity S, et al. PRRT2 mutations cause benign familial infantile epilepsy and infantile convulsions with choreoathetosis syndrome. American journal of human genetics 2012;90:152-160.
50. Gardiner AR, Bhatia KP, Stamelou M, et al. PRRT2 gene mutations: From paroxysmal dyskinesia to episodic ataxia and hemiplegic migraine. Neurology 2012.
51. Dale RC, Gardiner A, Antony J, Houlden H. Familial PRRT2 mutation with heterogeneous paroxysmal disorders including paroxysmal torticollis and hemiplegic migraine. Dev Med Child Neurol 2012;54:958-960.
52. Silveira-Moriyama L, Gardiner AR, Meyer E, et al. Clinical features of childhood-onset paroxysmal kinesigenic dyskinesia with PRRT2 gene mutations. Dev Med Child Neurol 2013.
53. Rossor AM, Polke JM, Houlden H, Reilly MM. Clinical implications of genetic advances in Charcot-Marie-Tooth disease. Nat Rev Neurol 2013;9:562-571.
54. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nature reviews Genetics 2013;14:295-300.
55. Sikkema-Raddatz B, Johansson LF, de Boer EN, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 2013;34:1035-1042.
56. Roach JC, Glusman G, Smit AF, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 2010;328:636-639.
57. Gibbs R, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H. The International HapMap Project. Nature 2003;426:789-796.
58. Simón-Sánchez J, Schulte C, Bras JM, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nature Genetics 2009;41:1308-1312.

59. Li T, Yang D, Zhong S, et al. Novel LRRK2 GTP-binding inhibitors reduced degeneration in Parkinson's disease cell and mouse models. Human Molecular Genetics 2014;23:6212-6222.
60. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet 2014.
61. Coon KD, Myers AJ, Craig DW, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry 2007;68:613-618.
62. Sawcer S, Hellenthal G, Pirinen M, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011;476:214-219.
63. Lee JC, Parkes M. Genome-wide association studies and Crohn's disease. Brief Funct Genomics 2011;10:71-76.
64. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385-389.
65. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature Genetics 2009;41:25-34.
66. Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet 2010;86:730-742.
67. Hamza TH, Zabetian CP, Tenesa A, et al. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nat Genet 2010;42:781-785.
68. Lee DW, Zhao X, Zhang F, Eisenberg E, Greene LE. Depletion of GAK/auxilin 2 inhibits receptor-mediated endocytosis and recruitment of both clathrin and clathrin adaptors. Journal of Cell Science 2005;118:4311-4321.
69. Dumitriu A, Pacheco CD, Wilk JB, et al. Cyclin-G-associated kinase modifies alpha-synuclein expression levels and toxicity in Parkinson's disease: results from the GenePD Study. Human Molecular Genetics 2011;20:1478-1487.
70. Ku CS, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: where are we now? Journal of Human Genetics 2010;55:195-206.
71. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 2006;63:168-174.
72. Kamboh MI, Demirci FY, Wang X, et al. Genome-wide association study of Alzheimer's disease. Transl Psychiatry 2012;2:e117.

73. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 2012;44:981-990.
74. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics 2010;42:1118-1125.
75. Tayebi N, Walker J, Stubblefield B, et al. Gaucher disease with parkinsonian manifestations: does glucocerebrosidase deficiency contribute to a vulnerability to parkinsonism? Molecular genetics and metabolism 2003;79:104-109.
76. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. American journal of human genetics 2014;95:5-23.
77. Lesage S, Brice A. Role of mendelian genes in "sporadic" Parkinson's disease. Parkinsonism & related disorders 2012;18 Suppl 1:S66-70.
78. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009;55:641-658.
79. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008;9:387-402.
80. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 2011;72:245-256.
81. Renton AE, Majounie E, Waite A, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 2011;72:257-268.
82. Cruts M, Gijselinck I, Van Langenhove T, van der Zee J, Van Broeckhoven C. Current insights into the C9orf72 repeat expansion diseases of the FTLD/ALS spectrum. Trends Neurosci 2013;36:450-459.
83. Coonrod EM, Margraf RL, Voelkerding KV. Translating exome sequencing from research to clinical diagnostics. Clin Chem Lab Med 2012;50:1161-1168.
84. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature Genetics 2010;42:790-793.
85. Quail MA, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. Nature methods 2008;5:1005-1010.
86. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature reviews Genetics 2014;15:121-132.
87. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res 2010;20:1165-1173.

88. Wang W, Wei Z, Lam TW, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci Rep 2011;1:55.
89. Hatem A, Bozdag D, Toland AE, Catalyurek UV. Benchmarking short sequence mapping tools. BMC Bioinformatics 2013;14:184.
90. Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 2012;28:2097-2105.
91. Derrien T, Estelle J, Marco Sola S, et al. Fast computation and applications of genome mappability. PLoS ONE 2012;7:e30377.
92. Knies K, Schuster B, Ameziane N, et al. Genotyping of fanconi anemia patients by whole exome sequencing: advantages and challenges. PLoS ONE 2012;7:e52648.
93. Li W, Freudenberg J, Miramontes P. Diminishing return for increased Mappability with longer sequencing reads: implications of the k-mer distributions in the human genome. BMC Bioinformatics 2014;15:2.
94. O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 2013;5:28.
95. Lyon GJ, Wang K. Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. Genome Med 2012;4:58.
96. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15:901-913.
97. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 2010;6:e1001025.
98. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 2010;20:110-121.
99. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 2005;15:1034-1050.
100. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863-874.
101. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012;7:e46688.
102. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248-249.
103. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;7:575-576.

104. Morton NE. Sequential tests for the detection of linkage. American journal of human genetics 1955;7:277-318.
105. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 2002;30:97-101.
106. Smith C. The Detection of Linkage in Human Genetics (with discussion). Journal of the Royal Statistical Society: Series B 1953;15:153-184.
107. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 1987;236:1567-1570.
108. Charlesworth G, Bhatia KP, Wood NW. The genetics of dystonia: new twists in an old tale. Brain : a journal of neurology 2013;136:2017-2037.
109. Fanh S, Marsden CD, B CD. Dystonia 2. Proceedings of the Second International Symposium on Torsion Dystonia. Harriman, New York, 1986. Adv Neurol 1988;50:1-705.
110. Defazio G, Berardelli A, Hallett M. Do primary adult-onset focal dystonias share aetiological factors? Brain : a journal of neurology 2007;130:1183-1193.
111. Breakefield XO, Blood AJ, Li Y, Hallett M, Hanson PI, Standaert DG. The pathophysiological basis of dystonias. Nature reviews Neuroscience 2008;9:222-234.
112. Geyer HL, Bressman SB. The diagnosis of dystonia. Lancet neurology 2006;5:780-790.
113. Paudel R, Hardy J, Revesz T, Holton JL, Houlden H. Review: Genetics and neuropathology of primary pure dystonia. Neuropathology and Applied Neurobiology 2012;38:520-534.
114. Zoons E, Dijkgraaf MG, Dijk JM, van Schaik IN, Tijssen MA. Botulinum toxin as treatment for focal dystonia: a systematic review of the pharmaco-therapeutic and pharmaco-economic value. Journal of neurology 2012.
115. Skogseid IM, Malt UF, Roislien J, Kerty E. Determinants and status of quality of life after long-term botulinum toxin therapy for cervical dystonia. European journal of neurology : the official journal of the European Federation of Neurological Societies 2007;14:1129-1137.
116. Soeder A, Kluger BM, Okun MS, et al. Mood and energy determinants of quality of life in dystonia. Journal of neurology 2009;256:996-1001.
117. Fahn S. Concept and classification of dystonia. Adv Neurol 1988;50:1-8.
118. Teo JT, van de Warrenburg BP, Schneider SA, Rothwell JC, Bhatia KP. Neurophysiological evidence for cerebellar dysfunction in primary focal dystonia. J Neurol Neurosurg Psychiatry 2009;80:80-83.
119. Quartarone A, Morgante F, Sant'angelo A, et al. Abnormal plasticity of sensorimotor circuits extends beyond the affected body part in focal dystonia. J Neurol Neurosurg Psychiatry 2008;79:985-990.

120. Argyelan M, Carbon M, Niethammer M, et al. Cerebellothalamocortical connectivity regulates penetrance in dystonia. The Journal of neuroscience : the official journal of the Society for Neuroscience 2009;29:9740-9747.
121. Albanese A, Asmus F, Bhatia KP, et al. EFNS guidelines on diagnosis and treatment of primary dystonias. European journal of neurology : the official journal of the European Federation of Neurological Societies 2011;18:5-18.
122. Kramer PL, de Leon D, Ozelius L, et al. Dystonia gene in Ashkenazi Jewish population is located on chromosome 9q32-34. Annals of neurology 1990;27:114-120.
123. Zlotogora J. Autosomal recessive, DYT2-like primary torsion dystonia: a new family. Neurology 2004;63:1340.
124. Khan NL, Wood NW, Bhatia KP. Autosomal recessive, DYT2-like primary torsion dystonia: a new family. Neurology 2003;61:1801-1803.
125. Quadri M, Federico A, Zhao T, et al. Mutations in SLC30A10 cause parkinsonism and dystonia with hypermanganesemia, polycythemia, and chronic liver disease. American journal of human genetics 2012;90:467-477.
126. Tuschl K, Clayton PT, Gospe SM, Jr., et al. Syndrome of hepatic cirrhosis, dystonia, polycythemia, and hypermanganesemia caused by mutations in SLC30A10, a manganese transporter in man. American journal of human genetics 2012;90:457-466.
127. Zanotti-Fregonara P, Vidailhet M, Kas A, et al. [123I]-FP-CIT and [99mTc]-HMPAO single photon emission computed tomography in a new sporadic case of rapid-onset dystonia-parkinsonism. J Neurol Sci 2008;273:148-151.
128. Romero-Lopez J, Moreno-Carretero MJ, Escriche-Jaime D, Corredera-Garcia E. RAPID-ONSET DYSTONIA-PARKINSONISM: SPORADIC FORM. Rev Neurologia 2008;47:638-640.
129. Waddy HM, Fletcher NA, Harding AE, Marsden CD. A genetic study of idiopathic focal dystonias. Annals of neurology 1991;29:320-324.
130. Schmidt A, Jabusch HC, Altenmuller E, et al. Etiology of musician's dystonia: familial or environmental? Neurology 2009;72:1248-1254.
131. Stojanovic M, Cvetkovic D, Kostic VS. A genetic study of idiopathic focal dystonias. Journal of neurology 1995;242:508-511.
132. Leube B, Kessler KR, Goecke T, Auburger G, Benecke R. Frequency of familial inheritance among 488 index patients with idiopathic focal dystonia and clinical variability in a large family. Movement disorders : official journal of the Movement Disorder Society 1997;12:1000-1006.

133. Kimmich O, Bradley D, Whelan R, et al. Sporadic adult onset primary torsion dystonia is a genetic disorder by the temporal discrimination test. Brain : a journal of neurology 2011;134:2656-2663.
134. Kramer PL, Heiman GA, Gasser T, et al. The DYT1 gene on 9q34 is responsible for most cases of early limb-onset idiopathic torsion dystonia in non-Jews. American journal of human genetics 1994;55:468-475.
135. Ozelius LJ, Kramer PL, de Leon D, et al. Strong allelic association between the torsion dystonia gene (DYT1) andloci on chromosome 9q34 in Ashkenazi Jews. American journal of human genetics 1992;50:619-628.
136. Ozelius LJ, Hewett JW, Page CE, et al. The early-onset torsion dystonia gene (DYT1) encodes an ATP-binding protein. Nature Genetics 1997;17:40-48.
137. Bressman S. Genetics of dystonia. J Neural Transm Suppl 2006:489-495.
138. Jamora RD, Tan EK, Liu CP, Kathirvel P, Burgunder JM, Tan LC. DYT1 mutations amongst adult primary dystonia patients in Singapore with review of literature comparing East and West. J Neurol Sci 2006;247:35-37.
139. Calakos N, Patel VD, Gottron M, et al. Functional evidence implicating a novel TOR1A mutation in idiopathic, late-onset focal dystonia. Journal of Medical Genetics 2010;47:646-650.
140. Zirn B, Grundmann K, Huppke P, et al. Novel TOR1A mutation p.Arg288Gln in early-onset dystonia (DYT1). J Neurol Neurosurg Psychiatry 2008;79:1327-1330.
141. Kabakci K, Hedrich K, Leung JC, et al. Mutations in DYT1: extension of the phenotypic and mutational spectrum. Neurology 2004;62:395-400.
142. Kamm C, Fischer H, Garavaglia B, et al. Susceptibility to DYT1 dystonia in European patients is modified by the D216H polymorphism. Neurology 2008;70:2261-2262.
143. Risch NJ, Bressman SB, Senthil G, Ozelius LJ. Intragenic Cis and Trans modification of genetic susceptibility in DYT1 torsion dystonia. American journal of human genetics 2007;80:1188-1193.
144. Kock N, Naismith TV, Boston HE, et al. Effects of genetic variations in the dystonia protein torsinA: identification of polymorphism at residue 216 as protein modifier. Human Molecular Genetics 2006;15:1355-1364.
145. Martino D, Gajos A, Gallo V, et al. Extragenetic factors and clinical penetrance of DYT1 dystonia: an exploratory study. Journal of neurology 2012.
146. Hjermind LE, Werdelin LM, Sorensen SA. Inherited and de novo mutations in sporadic cases of DYT1-dystonia. European journal of human genetics : EJHG 2002;10:213-216.

147. Konakova M, Huynh DP, Yong W, Pulst SM. Cellular distribution of torsin A and torsin B in normal human brain. Archives of neurology 2001;58:921-927.
148. Shashidharan P, Kramer BC, Walker RH, Olanow CW, Brin MF. Immunohistochemical localization and distribution of torsinA in normal human and rat brain. Brain research 2000;853:197-206.
149. Rostasy K, Augood SJ, Hewett JW, et al. TorsinA protein and neuropathology in early onset generalized dystonia with GAG deletion. Neurobiology of Disease 2003;12:11-24.
150. Kakazu Y, Koh JY, Ho KW, Gonzalez-Alegre P, Harata NC. Synaptic vesicle recycling is enhanced by torsinA that harbors the DYT1 dystonia mutation. Synapse 2012;66:453-464.
151. Henriksen C, Madsen LB, Bendixen C, Larsen K. Characterization of the porcine TOR1A gene: The first step towards generation of a pig model for dystonia. Gene 2009;430:105-115.
152. Goodchild RE, Dauer WT. The AAA+ protein torsinA interacts with a conserved domain present in LAP1 and a novel ER protein. The Journal of cell biology 2005;168:855-862.
153. Goodchild RE, Kim CE, Dauer WT. Loss of the dystonia-associated protein torsinA selectively disrupts the neuronal nuclear envelope. Neuron 2005;48:923-932.
154. Hewett JW, Tannous B, Niland BP, et al. Mutant torsinA interferes with protein processing through the secretory pathway in DYT1 dystonia cells. Proceedings of the National Academy of Sciences of the United States of America 2007;104:7271-7276.
155. Hewett JW, Zeng J, Niland BP, Bragg DC, Breakefield XO. Dystonia-causing mutant torsinA inhibits cell adhesion and neurite extension through interference with cytoskeletal dynamics. Neurobiology of Disease 2006;22:98-111.
156. Almasy L, Bressman SB, Raymond D, et al. Idiopathic torsion dystonia linked to chromosome 8 in two Mennonite families. Annals of neurology 1997;42:670-673.
157. Fuchs T, Gavarini S, Saunders-Pullman R, et al. Mutations in the THAP1 gene are responsible for DYT6 primary torsion dystonia. Nature Genetics 2009;41:286-288.
158. Saunders-Pullman R, Raymond D, Senthil G, et al. Narrowing the DYT6 dystonia region and evidence for locus heterogeneity in the Amish-Mennonites. Am J Med Genet A 2007;143A:2098-2105.
159. Houlden H, Schneider SA, Paudel R, et al. THAP1 mutations (DYT6) are an additional cause of early-onset dystonia. Neurology 2010;74:846-850.
160. Xiromerisiou G, Houlden H, Scarmeas N, et al. THAP1 mutations and dystonia phenotypes: Genotype phenotype correlations. Movement disorders : official journal of the Movement Disorder Society 2012;27:1290-1294.

161. Roussigne M, Cayrol C, Clouaire T, Amalric F, Girard JP. THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. Oncogene 2003;22:2432-2442.
162. Clouaire T, Roussigne M, Ecochard V, Mathe C, Amalric F, Girard JP. The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. Proceedings of the National Academy of Sciences of the United States of America 2005;102:6907-6912.
163. Duan W, Zhang Z, Gash DM, Mattson MP. Participation of prostate apoptosis response-4 in degeneration of dopaminergic neurons in models of Parkinson's disease. Annals of neurology 1999;46:587-597.
164. Guo Q, Fu W, Xie J, et al. Par-4 is a mediator of neuronal degeneration associated with the pathogenesis of Alzheimer disease. Nature Medicine 1998;4:957-962.
165. Pedersen WA, Luo H, Kruman I, Kasarskis E, Mattson MP. The prostate apoptosis response-4 protein participates in motor neuron degeneration in amyotrophic lateral sclerosis. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 2000;14:913-924.
166. Kaiser FJ, Osmanoric A, Rakovic A, et al. The dystonia gene DYT1 is repressed by the transcription factor THAP1 (DYT6). Annals of neurology 2010;68:554-559.
167. Gavarini S, Cayrol C, Fuchs T, et al. Direct interaction between causative genes of DYT1 and DYT6 primary dystonia. Annals of neurology 2010;68:549-553.
168. Kaiser FJ, Bassler N, Jakel O. COTS Silicon diodes as radiation detectors in proton and heavy charged particle radiotherapy 1. Radiat Environ Biophys 2010;49:365-371.
169. Bragg DC, Armata IA, Nery FC, Breakefield XO, Sharma N. Molecular pathways in dystonia. Neurobiology of Disease 2011;42:136-147.
170. Grundmann K, Reischmann B, Vanhoutte G, et al. Overexpression of human wildtype torsinA and human DeltaGAG torsinA in a transgenic mouse model causes phenotypic abnormalities. Neurobiology of Disease 2007;27:190-206.
171. Fuchs T, Saunders-Pullman R, Masuho I, et al. Mutations in GNAL cause primary torsion dystonia. Nature Genetics 2012;45:88-92.
172. Herve D, Levi-Strauss M, Marey-Semper I, et al. G(olf) and Gs in rat basal ganglia: possible involvement of G(olf) in the coupling of dopamine D1 receptor with adenylyl cyclase. The Journal of neuroscience : the official journal of the Society for Neuroscience 1993;13:2237-2248.

173. Zhuang X, Belluscio L, Hen R. G(olf)alpha mediates dopamine D1 receptor signaling. The Journal of neuroscience : the official journal of the Society for Neuroscience 2000;20:RC91.
174. Corvol JC, Studler JM, Schonn JS, Girault JA, Herve D. Galpha(olf) is necessary for coupling D1 and A2a receptors to adenylyl cyclase in the striatum. Journal of Neurochemistry 2001;76:1585-1588.
175. Leube B, Hendgen T, Kessler KR, Knapp M, Benecke R, Auburger G. Evidence for DYT7 being a common cause of cervical dystonia (torticollis) in Central Europe. Am J Med Genet 1997;74:529-532.
176. Berrettini WH. Susceptibility loci for bipolar disorder: overlap with inherited vulnerability to schizophrenia. Biol Psychiatry 2000;47:245-251.
177. Laurin N, Ickowicz A, Pathare T, et al. Investigation of the G protein subunit Galphaolf gene (GNAL) in attention deficit/hyperactivity disorder. J Psychiatr Res 2008;42:117-124.
178. Segurado R, Detera-Wadleigh SD, Levinson DF, et al. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. American journal of human genetics 2003;73:49-62.
179. Winter P, Kamm C, Biskup S, et al. DYT7 gene locus for cervical dystonia on chromosome 18p is questionable. Movement disorders : official journal of the Movement Disorder Society 2012;27:1819-1821.
180. Belluscio L, Gold GH, Nemes A, Axel R. Mice deficient in G(olf) are anosmic. Neuron 1998;20:69-81.
181. Xiao J, Uitti RJ, Zhao Y, et al. Mutations in CIZ1 cause adult onset primary cervical dystonia. Annals of neurology 2012;71:458-469.
182. Uitti RJ, Maraganore DM. Adult onset familial cervical dystonia: report of a family including monozygotic twins. Movement disorders : official journal of the Movement Disorder Society 1993;8:489-494.
183. Klein C, König IR, Lohmann K. Exome sequencing for gene discovery: Time to set standard criteria. Annals of neurology 2012:627-628.
184. Parker N. Hereditary whispering dysphonia. J Neurol Neurosurg Psychiatry 1985;48:218-224.
185. Wilcox RA, Winkler S, Lohmann K, Klein C. Whispering dysphonia in an Australian family (DYT4): a clinical and genetic reappraisal. Movement disorders : official journal of the Movement Disorder Society 2011;26:2404-2408.

186. Hersheson J, Mencacci NE, Davis M, et al. Mutations in the autoregulatory domain of β-tubulin 4a cause hereditary dystonia. Annals of neurology 2012:n/a-n/a.
187. Lohmann K, Wilcox RA, Winkler S, et al. Whispering dysphonia (DYT4 dystonia) is caused by a mutation in the TUBB4 gene. Annals of neurology 2012:n/a-n/a.
188. Yen TJ, Gay DA, Pachter JS, Cleveland DW. Autoregulated changes in stability of polyribosome-bound beta-tubulin mRNAs are specified by the first 13 translated nucleotides. Molecular and cellular biology 1988;8:1224-1235.
189. Yen TJ, Machlin PS, Cleveland DW. Autoregulated instability of beta-tubulin mRNAs by recognition of the nascent amino terminus of beta-tubulin. Nature 1988;334:580-585.
190. Bruno MK, Lee HY, Auburger GW, et al. Genotype-phenotype correlation of paroxysmal nonkinesigenic dyskinesia. Neurology 2007;68:1782-1789.
191. Lee HY, Xu Y, Huang Y, et al. The gene for paroxysmal non-kinesigenic dyskinesia encodes an enzyme in a stress response pathway. Human Molecular Genetics 2004;13:3161-3170.
192. Djarmati A, Svetel M, Momcilovic D, Kostic V, Klein C. Significance of recurrent mutations in the myofibrillogenesis regulator 1 gene. Archives of neurology 2005;62:1641.
193. Rainier S, Thomas D, Tokarz D, et al. Myofibrillogenesis regulator 1 gene mutations cause paroxysmal dystonic choreoathetosis. Archives of neurology 2004;61:1025-1029.
194. Chen DH, Matsushita M, Rainier S, et al. Presence of alanine-to-valine substitutions in myofibrillogenesis regulator 1 in paroxysmal nonkinesigenic dyskinesia: confirmation in 2 kindreds. Archives of neurology 2005;62:597-600.
195. Hempelmann A, Kumar S, Muralitharan S, Sander T. Myofibrillogenesis regulator 1 gene (MR-1) mutation in an Omani family with paroxysmal nonkinesigenic dyskinesia. Neuroscience letters 2006;402:118-120.
196. Ghezzi D, Viscomi C, Ferlini A, et al. Paroxysmal non-kinesigenic dyskinesia is caused by mutations of the MR-1 mitochondrial targeting sequence. Human Molecular Genetics 2009;18:1058-1064.
197. Shen Y, Lee HY, Rawson J, et al. Mutations in PNKD causing paroxysmal dyskinesia alters protein cleavage and stability. Human Molecular Genetics 2011;20:2322-2332.
198. Bruno MK, Hallett M, Gwinn-Hardy K, et al. Clinical evaluation of idiopathic paroxysmal kinesigenic dyskinesia: new diagnostic criteria. Neurology 2004;63:2280-2287.
199. Smith LA, Heersema PH. Periodic Dystonia. Mayo Clinic Proceedings 1941:4.
200. Muller U, Steinberger D, Nemeth AH. Clinical and molecular genetics of primary dystonias. Neurogenetics 1998;1:165-177.

201. Hedera P, Xiao J, Puschmann A, Momcilovic D, Wu SW, Ledoux MS. Novel PRRT2 mutation in an African-American family with paroxysmal kinesigenic dyskinesia. BMC Neurol 2012;12:93.
202. Friedman J, Olvera J, Silhavy JL, Gabriel SB, Gleeson JG. Mild paroxysmal kinesigenic dyskinesia caused by PRRT2 missense mutation with reduced penetrance. Neurology 2012;79:946-948.
203. Steinlein OK, Villain M, Korenke C. The PRRT2 mutation c.649dupC is the so far most frequent cause of benign familial infantile convulsions. Seizure 2012;21:740-742.
204. Cao L, Huang XJ, Zheng L, Xiao Q, Wang XJ, Chen SD. Identification of a novel PRRT2 mutation in patients with paroxysmal kinesigenic dyskinesias and c.649dupC as a mutation hot-spot. Parkinsonism & related disorders 2012;18:704-706.
205. Stelzl U, Worm U, Lalowski M, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005;122:957-968.
206. Li J, Zhu X, Wang X, et al. Targeted genomic sequencing identifies PRRT2 mutations as a cause of paroxysmal kinesigenic choreoathetosis. Journal of Medical Genetics 2012;49:76-78.
207. Suls A, Dedeken P, Goffin K, et al. Paroxysmal exercise-induced dyskinesia and epilepsy is due to mutations in SLC2A1, encoding the glucose transporter GLUT1. Brain : a journal of neurology 2008;131:1831-1844.
208. Weber YG, Storch A, Wuttke TV, et al. GLUT1 mutations are a cause of paroxysmal exertion-induced dyskinesias and induce hemolytic anemia by a cation leak. J Clin Invest 2008;118:2157-2168.
209. Perez-Duenas B, Prior C, Ma Q, et al. Childhood chorea with cerebral hypotrophy: a treatable GLUT1 energy failure syndrome. Archives of neurology 2009;66:1410-1414.
210. Marin-Valencia I, Good LB, Ma Q, et al. Glut1 deficiency (G1D): epilepsy and metabolic dysfunction in a mouse model of the most common human phenotype. Neurobiology of Disease 2012;48:92-101.
211. Magistretti PJ, Pellerin L. Cellular mechanisms of brain energy metabolism. Relevance to functional brain imaging and to neurodegenerative disorders. Annals of the New York Academy of Sciences 1996;777:380-387.
212. Leen WG, Klepper J, Verbeek MM, et al. Glucose transporter-1 deficiency syndrome: the expanding clinical and genetic spectrum of a treatable disorder. Brain : a journal of neurology 2010;133:655-670.
213. Brockmann K. The expanding phenotype of GLUT1-deficiency syndrome. Brain and Development 2009;31:545-552.

214. Asmus F, Devlin A, Munz M, Zimprich A, Gasser T, Chinnery PF. Clinical differentiation of genetically proven benign hereditary chorea and myoclonus-dystonia. Movement disorders : official journal of the Movement Disorder Society 2007;22:2104-2109.
215. Asmus F, Gasser T. Inherited myoclonus-dystonia. Adv Neurol 2004;94:113-119.
216. Nardocci N. Myoclonus-dystonia syndrome. Handb Clin Neurol 2011;100:563-575.
217. Kinugawa K, Vidailhet M, Clot F, Apartis E, Grabli D, Roze E. Myoclonus-dystonia: an update. Movement disorders : official journal of the Movement Disorder Society 2009;24:479-489.
218. Esapa CT, Waite A, Locke M, et al. SGCE missense mutations that cause myoclonus-dystonia syndrome impair epsilon-sarcoglycan trafficking to the plasma membrane: modulation by ubiquitination and torsinA. Human Molecular Genetics 2007;16:327-342.
219. Grabowski M, Zimprich A, Lorenz-Depiereux B, et al. The epsilon-sarcoglycan gene (SGCE), mutated in myoclonus-dystonia syndrome, is maternally imprinted. European journal of human genetics : EJHG 2003;11:138-144.
220. Hess CW, Raymond D, Aguiar Pde C, et al. Myoclonus-dystonia, obsessive-compulsive disorder, and alcohol dependence in SGCE mutation carriers. Neurology 2007;68:522-524.
221. Peall KJ, Smith DJ, Kurian MA, et al. SGCE mutations cause psychiatric disorders: clinical and genetic characterization. Brain : a journal of neurology 2013;136:294-303.
222. Roze E, Apartis E, Clot F, et al. Myoclonus-dystonia: clinical and electrophysiologic pattern related to SGCE mutations. Neurology 2008;70:1010-1016.
223. Luciano MS, Ozelius L, Sims K, Raymond D, Liu L, Saunders-Pullman R. Responsiveness to levodopa in epsilon-sarcoglycan deletions. Movement disorders : official journal of the Movement Disorder Society 2009;24:425-428.
224. de Carvalho Aguiar P, Sweadner KJ, Penniston JT, et al. Mutations in the Na+/K+-ATPase Œ±3 Gene ATP1A3 Are Associated with Rapid-Onset Dystonia Parkinsonism. Neuron 2004;43:169-175.
225. Rodacker V, Toustrup-Jensen M, Vilsen B. Mutations Phe785Leu and Thr618Met in Na+,K+-ATPase, associated with familial rapid-onset dystonia parkinsonism, interfere with Na+ interaction by distinct mechanisms. The Journal of biological chemistry 2006;281:18539-18548.
226. Blanco-Arias P, Einholm AP, Mamsa H, et al. A C-terminal mutation of ATP1A3 underscores the crucial role of sodium affinity in the pathophysiology of rapid-onset dystonia-parkinsonism. Human Molecular Genetics 2009;18:2370-2377.

227. Brashear A, Dobyns WB, de Carvalho Aguiar P, et al. The phenotypic spectrum of rapid-onset dystonia-parkinsonism (RDP) and mutations in the ATP1A3 gene. Brain : a journal of neurology 2007;130:828-835.
228. Brashear A, Mulholland GK, Zheng QH, Farlow MR, Siemers ER, Hutchins GD. PET imaging of the pre-synaptic dopamine uptake sites in rapid-onset dystonia-parkinsonism (RDP). Movement disorders : official journal of the Movement Disorder Society 1999;14:132-137.
229. Deutschlander A, Asmus F, Gasser T, Steude U, Botzel K. Sporadic rapid-onset dystonia-parkinsonism syndrome: failure of bilateral pallidal stimulation. Movement disorders : official journal of the Movement Disorder Society 2005;20:254-257.
230. Kamm C, Fogel W, Wachter T, et al. Novel ATP1A3 mutation in a sporadic RDP patient with minimal benefit from deep brain stimulation. Neurology 2008;70:1501-1503.
231. Heinzen EL, Swoboda KJ, Hitomi Y, et al. De novo mutations in ATP1A3 cause alternating hemiplegia of childhood. Nature Genetics 2012;44:1030-1034.
232. Camargos S, Scholz S, Simon-Sanchez J, et al. DYT16, a novel young-onset dystonia-parkinsonism disorder: identification of a segregating mutation in the stress-response protein PRKRA. Lancet neurology 2008;7:207-215.
233. Patel CV, Handy I, Goldsmith T, Patel RC. PACT, a stress-modulated cellular activator of interferon-induced double-stranded RNA-activated protein kinase, PKR. The Journal of biological chemistry 2000;275:37993-37998.
234. Brautigam C, Wevers RA, Jansen RJ, et al. Biochemical hallmarks of tyrosine hydroxylase deficiency. Clinical chemistry 1998;44:1897-1904.
235. Ichinose H, Ohye T, Takahashi E, et al. Hereditary progressive dystonia with marked diurnal fluctuation caused by mutations in the GTP cyclohydrolase I gene. Nature Genetics 1994;8:236-242.
236. Bonafe L, Thony B, Penzien JM, Czarnecki B, Blau N. Mutations in the sepiapterin reductase gene cause a novel tetrahydrobiopterin-dependent monoamine-neurotransmitter deficiency without hyperphenylalaninemia. American journal of human genetics 2001;69:269-277.
237. Furukawa Y. Update on dopa-responsive dystonia: locus heterogeneity and biochemical features. Adv Neurol 2004;94:127-138.
238. Ichinose H, Nagatsu T. Molecular genetics of hereditary dystonia--mutations in the GTP cyclohydrolase I gene. Brain Res Bull 1997;43:35-38.
239. Hwu WL, Chiou YW, Lai SY, Lee YM. Dopa-responsive dystonia is induced by a dominant-negative mechanism. Annals of neurology 2000;48:609-613.

240. Bandmann O, Goertz M, Zschocke J, et al. The phenylalanine loading test in the differential diagnosis of dystonia. Neurology 2003;60:700-702.
241. Furukawa Y. Genetics and biochemistry of dopa-responsive dystonia: significance of striatal tyrosine hydroxylase protein loss. Adv Neurol 2003;91:401-410.
242. Nygaard TG, Trugman JM, de Yebenes JG, Fahn S. Dopa-responsive dystonia: the spectrum of clinical manifestations in a large North American family. Neurology 1990;40:66-69.
243. Segawa M, Nomura Y, Nishiyama N. Autosomal dominant guanosine triphosphate cyclohydrolase I deficiency (Segawa disease). Annals of neurology 2003;54 Suppl 6:S32-45.
244. Cobb SA, Wider C, Ross OA, et al. GCH1 in early-onset Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2009;24:2070-2075.
245. Trender-Gerhard I, Sweeney MG, Schwingenschuh P, et al. Autosomal-dominant GTPCH1-deficient DRD: clinical characteristics and long-term outcome of 34 patients. J Neurol Neurosurg Psychiatry 2009;80:839-845.
246. Tassin J, Durr A, Bonnet AM, et al. Levodopa-responsive dystonia. GTP cyclohydrolase I or parkin mutations? Brain : a journal of neurology 2000;123 (Pt 6):1112-1121.
247. Stamelou M, Mencacci NE, Cordivari C, et al. Myoclonus-dystonia syndrome due to tyrosine hydroxylase deficiency. Neurology 2012;79:435-441.
248. Van Hove JL, Steyaert J, Matthijs G, et al. Expanded motor and psychiatric phenotype in autosomal dominant Segawa syndrome due to GTP cyclohydrolase deficiency. J Neurol Neurosurg Psychiatry 2006;77:18-23.
249. Clot F, Grabli D, Cazeneuve C, et al. Exhaustive analysis of BH4 and dopamine biosynthesis genes in patients with Dopa-responsive dystonia. Brain : a journal of neurology 2009;132:1753-1763.
250. Ichinose H, Ohye T, Matsuda Y, et al. Characterization of mouse and human GTP cyclohydrolase I genes. Mutations in patients with GTP cyclohydrolase I deficiency. The Journal of biological chemistry 1995;270:10062-10071.
251. Hwu WL, Wang PJ, Hsiao KJ, Wang TR, Chiou YW, Lee YM. Dopa-responsive dystonia induced by a recessive GTP cyclohydrolase I mutation. Human genetics 1999;105:226-230.
252. Nardocci N, Zorzi G, Blau N, et al. Neonatal dopa-responsive extrapyramidal syndrome in twins with recessive GTPCH deficiency. Neurology 2003;60:335-337.
253. Horvath GA, Stockler-Ipsiroglu SG, Salvarinova-Zivkovic R, et al. Autosomal recessive GTP cyclohydrolase I deficiency without hyperphenylalaninemia: evidence of a phenotypic

continuum between dominant and recessive forms. Molecular genetics and metabolism 2008;94:127-131.

254. Opladen T, Hoffmann G, Horster F, et al. Clinical and biochemical characterization of patients with early infantile onset of autosomal recessive GTP cyclohydrolase I deficiency without hyperphenylalaninemia. Movement disorders : official journal of the Movement Disorder Society 2011;26:157-161.

255. Ludecke B, Knappskog PM, Clayton PT, et al. Recessively inherited L-DOPA-responsive parkinsonism in infancy caused by a point mutation (L205P) in the tyrosine hydroxylase gene. Human Molecular Genetics 1996;5:1023-1028.

256. Willemsen MA, Verbeek MM, Kamsteeg EJ, et al. Tyrosine hydroxylase deficiency: a treatable disorder of brain catecholamine biosynthesis. Brain : a journal of neurology 2010;133:1810-1822.

257. Zhou QY, Quaife CJ, Palmiter RD. Targeted disruption of the tyrosine hydroxylase gene reveals that catecholamines are required for mouse fetal development. Nature 1995;374:640-643.

258. Blau N, Bonafe L, Thony B. Tetrahydrobiopterin deficiencies without hyperphenylalaninemia: diagnosis and genetics of dopa-responsive dystonia and sepiapterin reductase deficiency. Molecular genetics and metabolism 2001;74:172-185.

259. Jones CL, Vasquez-Vivar J, Kalyanaraman B, Griscavage-Ennis M, Gross SS. Tetrahydropterins but not dihydropterins attenuate the reduction of superoxide from eNOS. Pteridines 2001:52-53.

260. Friedman J, Roze E, Abdenur JE, et al. Sepiapterin reductase deficiency: a treatable mimic of cerebral palsy. Annals of neurology 2012;71:520-530.

261. Bonafe L, Thony B, Leimbacher W, Kierat L, Blau N. Diagnosis of dopa-responsive dystonia and other tetrahydrobiopterin disorders by the study of biopterin metabolism in fibroblasts. Clinical chemistry 2001;47:477-485.

262. Song CH, Fan X, Exeter CJ, Hess EJ, Jinnah HA. Functional analysis of dopaminergic systems in a DYT1 knock-in mouse model of dystonia. Neurobiology of Disease 2012;48:66-78.

263. Coverley D, Marr J, Ainscough J. Ciz1 promotes mammalian DNA replication. Journal of Cell Science 2005;118:101-112.

264. Atai NA, Ryan SD, Kothary R, Breakefield XO, Nery FC. Untethering the nuclear envelope and cytoskeleton: biologically distinct dystonias arising from a common cellular dysfunction. Int J Cell Biol 2012;2012:634214.

265. Anastasi G, Tomasello F, Di Mauro D, et al. Expression of sarcoglycans in the human cerebral cortex: an immunohistochemical and molecular study. Cells Tissues Organs 2012;196:470-480.
266. Warner TT, Granata A, Schiavo G. TorsinA and DYT1 dystonia: a synaptopathy? Biochem Soc Trans 2010;38:452-456.
267. Makino S, Kaji R, Ando S, et al. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. American journal of human genetics 2007;80:393-406.
268. Müller U. The monogenic primary dystonias. Brain 2009;132:2005-2025.
269. Waters CH, Faust PL, Powers J, et al. Neuropathology of lubag (x-linked dystonia parkinsonism). Movement disorders : official journal of the Movement Disorder Society 1993;8:387-390.
270. Evidente VG, Advincula J, Esteban R, et al. Phenomenology of "Lubag" or X-linked dystonia-parkinsonism. Mov Disord 2002;17:1271-1277.
271. Kasperaviciute D, Catarino CB, Heinzen EL, et al. Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. Brain : a journal of neurology 2010;133:2136-2147.
272. Clarimon J, Asgeirsson H, Singleton A, et al. Torsin A haplotype predisposes to idiopathic dystonia. Annals of neurology 2005;57:765-767.
273. Kamm C, Asmus F, Mueller J, et al. Strong genetic evidence for association of TOR1A/TOR1B with idiopathic dystonia. Neurology 2006;67:1857-1859.
274. Sharma N, Franco RA, Jr., Kuster JK, et al. Genetic evidence for an association of the TOR1A locus with segmental/focal dystonia. Movement disorders : official journal of the Movement Disorder Society 2010;25:2183-2187.
275. Bruggemann N, Kock N, Lohmann K, et al. The D216H variant in the DYT1 gene: a susceptibility factor for dystonia in familial cases? Neurology 2009;72:1441-1443.
276. Clarimon J, Brancati F, Peckham E, et al. Assessing the role of DRD5 and DYT1 in two different case-control series with primary blepharospasm. Movement disorders : official journal of the Movement Disorder Society 2007;22:162-166.
277. Naiya T, Biswas A, Neogi R, et al. Clinical characterization and evaluation of DYT1 gene in Indian primary dystonia patients. Acta Neurol Scand 2006;114:210-215.
278. Hague S, Klaffke S, Clarimon J, et al. Lack of association with TorsinA haplotype in German patients with sporadic dystonia. Neurology 2006;66:951-952.
279. Sibbing D, Asmus F, Konig IR, et al. Candidate gene studies in focal dystonia. Neurology 2003;61:1097-1101.

280. Kaffe M, Gross N, Castrop F, et al. Mutational screening of THAP1 in a German population with primary dystonia. Parkinsonism & related disorders 2012;18:104-106.
281. Xiao J, Zhao Y, Bastian RW, et al. The c.-237_236GA>TT THAP1 sequence variant does not increase risk for primary dystonia. Movement disorders : official journal of the Movement Disorder Society 2011;26:549-552.
282. Newman JR, Sutherland GT, Boyle RS, et al. Common polymorphisms in dystonia-linked genes and susceptibility to the sporadic primary dystonias. Parkinsonism & related disorders 2012;18:351-357.
283. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2011;13:101.
284. Lorenz MG, Cortes LM, Lorenz JJ, Liu ET. Strategy for the design of custom cDNA microarrays. Biotechniques 2003;34:1264-1270.
285. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. SNP detection for massively parallel whole-genome resequencing. Genome Research 2009;19:1124-1132.
286. Millar T, Walker R, Arango JC, et al. Tissue and organ donation for research in forensic pathology: the MRC Sudden Death Brain and Tissue Bank. J Pathol 2007;213:369-375.
287. Beach TG, Sue LI, Walker DG, et al. The Sun Health Research Institute Brain Donation Program: description and experience, 1987-2007. Cell Tissue Bank 2008;9:229-245.
288. Trabzuni D, Ryten M, Walker R, et al. Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. Journal of Neurochemistry 2011;119:275-282.
289. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 2003;31:e15.
290. Trabzuni D, Wray S, Vandrovcova J, et al. MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies. Human Molecular Genetics 2012.
291. Velickovic M, Benabou R, Brin MF. Cervical dystonia pathophysiology and treatment options. Drugs 2001;61:1921-1943.
292. EDSE ESoDiECG. A prevalence study of primary dystonia in eight European countries. Journal of neurology 2000;247:787-792.
293. Duffey PO, Butler AG, Hawthorne MR, Barnes MP. The epidemiology of the primary dystonias in the north of England. Adv Neurol 1998;78:121-125.
294. Chan J, Brin MF, Fahn S. Idiopathic cervical dystonia: clinical characteristics. Movement disorders : official journal of the Movement Disorder Society 1991;6:119-126.

295. Defazio G, Aniello MS, Masi G, Lucchese V, De Candia D, Martino D. Frequency of familial aggregation in primary adult-onset cranial cervical dystonia. Neurol Sci 2003;24:168-169.

296. Munchau A, Valente EM, Davis MB, et al. A Yorkshire family with adult-onset cranio-cervical primary torsion dystonia. Movement disorders : official journal of the Movement Disorder Society 2000;15:954-959.

297. Vaarmann A, Gandhi S, Gourine AV, Abramov AY. Novel pathway for an old neurotransmitter: dopamine-induced neuronal calcium signalling via receptor-independent mechanisms. Cell Calcium 2010;48:176-182.

298. Oyarzabal A, Martinez-Pardo M, Merinero B, et al. A novel regulatory defect in the branched-chain alpha-keto acid dehydrogenase complex due to a mutation in the PPM1K gene causes a mild variant phenotype of maple syrup urine disease. Human mutation 2013;34:355-362.

299. Young RP, Hopkins RJ, Hay BA, Whittington CF, Epton MJ, Gamble GD. FAM13A locus in COPD is independently associated with lung cancer - evidence of a molecular genetic link between COPD and lung cancer. Appl Clin Genet 2011;4:1-10.

300. Cho MH, Boutaoui N, Klanderman BJ, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nature Genetics 2010;42:200-202.

301. Nakashima A, Yoshino K, Miyamoto T, et al. Identification of TBC7 having TBC domain as a novel binding protein to TSC1-TSC2 complex. Biochem Biophys Res Commun 2007;361:218-223.

302. Alfaiz AA, Micale L, Mandriani B, et al. TBC1D7 mutations are associated with intellectual disability, macrocrania, patellar dislocation, and celiac disease. Human mutation 2014;35:447-451.

303. Capo-Chichi JM, Tcherkezian J, Hamdan FF, et al. Disruption of TBC1D7, a subunit of the TSC1-TSC2 protein complex, in intellectual disability and megalencephaly. Journal of Medical Genetics 2013;50:740-744.

304. Quinn NP, Schneider SA, Schwingenschuh P, Bhatia KP. Tremor—some controversial aspects. Movement Disorders 2011;26:18-23.

305. Kang HJ, Kawasawa YI, Cheng F, et al. Spatio-temporal transcriptome of the human brain. Nature 2011;478:483-489.

306. Johnson MB, Kawasawa YI, Mason CE, et al. Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. Neuron 2009;62:494-509.

307. Duran C, Qu Z, Osunkoya AO, Cui Y, Hartzell HC. ANOs 3-7 in the anoctamin/Tmem16 Cl- channel family are intracellular proteins. Am J Physiol Cell Physiol 2012;302:C482-493.

308. Milenkovic VM, Brockmann M, Stohr H, Weber BH, Strauss O. Evolution and functional divergence of the anoctamin family of membrane proteins. BMC Evol Biol 2010;10:319.

309. Gritli-Linde A, Vaziri Sani F, Rock JR, et al. Expression patterns of the Tmem16 gene family during cephalic development in the mouse. Gene Expr Patterns 2009;9:178-191.

310. Duran C, Hartzell HC. Physiological roles and diseases of Tmem16/Anoctamin proteins: are they all chloride channels? Acta Pharmacol Sin 2011;32:685-692.

311. Duvvuri U, Shiwarski DJ, Xiao D, et al. TMEM16A Induces MAPK and Contributes Directly to Tumorigenesis and Cancer Progression. Cancer Res 2012;72:3270-3281.

312. Bolduc V, Marlow G, Boycott KM, et al. Recessive mutations in the putative calcium-activated chloride channel Anoctamin 5 cause proximal LGMD2L and distal MMD3 muscular dystrophies. American journal of human genetics 2010;86:213-221.

313. Penttila S, Palmio J, Suominen T, et al. Eight new mutations and the expanding phenotype variability in muscular dystrophy caused by ANO5. Neurology 2012;78:897-903.

314. Vermeer S, Hoischen A, Meijer RP, et al. Targeted next-generation sequencing of a 12.5 Mb homozygous region reveals ANO10 mutations in patients with autosomal-recessive cerebellar ataxia. American journal of human genetics 2010;87:813-819.

315. Suzuki J, Umeda M, Sims PJ, Nagata S. Calcium-dependent phospholipid scrambling by TMEM16F. Nature 2010;468:834-838.

316. Caputo A, Caci E, Ferrera L, et al. TMEM16A, a membrane protein associated with calcium-dependent chloride channel activity. Science 2008;322:590-594.

317. Huang WC, Xiao S, Huang F, Harfe BD, Jan YN, Jan LY. Calcium-activated chloride channels (CaCCs) regulate action potential and synaptic response in hippocampal neurons. Neuron 2012;74:179-192.

318. Hartzell C, Putzier I, Arreola J. Calcium-activated chloride channels. Annu Rev Physiol 2005;67:719-758.

319. Cenedese V, Betto G, Celsi F, Cherian OL, Pifferi S, Menini A. The voltage dependence of the TMEM16B/anoctamin2 calcium-activated chloride channel is modified by mutations in the first putative intracellular loop. J Gen Physiol 2012;139:285-294.

320. Yang YD, Cho H, Koo JY, et al. TMEM16A confers receptor-activated calcium-dependent chloride conductance. Nature 2008;455:1210-1215.

321. Yu K, Duran C, Qu Z, Cui YY, Hartzell HC. Explaining calcium-dependent gating of anoctamin-1 chloride channels requires a revised topology. Circ Res 2012;110:990-999.
322. Gong HC, Hang J, Kohler W, Li L, Su TZ. Tissue-specific expression and gabapentin-binding properties of calcium channel alpha2delta subunit subtypes. J Membr Biol 2001;184:35-43.
323. Hanke S, Bugert P, Chudek J, Kovacs G. Cloning a calcium channel alpha2delta-3 subunit gene from a putative tumor suppressor gene region at chromosome 3p21.1 in conventional renal cell carcinoma. Gene 2001;264:69-75.
324. Leone PE, Gonzalez MB, Elosua C, et al. Integration of global spectral karyotyping, CGH arrays, and expression arrays reveals important genes in the pathogenesis of glioblastoma multiforme. Ann Surg Oncol 2012;19:2367-2379.
325. Palmieri C, Rudraraju B, Monteverde M, et al. Methylation of the calcium channel regulatory subunit alpha2delta-3 (CACNA2D3) predicts site-specific relapse in oestrogen receptor-positive primary breast carcinomas. Br J Cancer 2012;107:375-381.
326. Wanajo A, Sasaki A, Nagasaki H, et al. Methylation of the calcium channel-related gene, CACNA2D3, is frequent and a poor prognostic factor in gastric cancer. Gastroenterology 2008;135:580-590.
327. Li Y, Zhu CL, Nie CJ, et al. Investigation of tumor suppressing function of CACNA2D3 in esophageal squamous cell carcinoma. PLoS ONE 2013;8:e60027.
328. Hoppa MB, Lana B, Margas W, Dolphin AC, Ryan TA. alpha2delta expression sets presynaptic calcium channel abundance and release probability. Nature 2012;486:122-125.
329. Davies A, Kadurin I, Alvarez-Laviada A, et al. The alpha2delta subunits of voltage-gated calcium channels form GPI-anchored proteins, a posttranslational modification essential for function. Proc Natl Acad Sci U S A 2010;107:1654-1659.
330. Neely GG, Hess A, Costigan M, et al. A genome-wide Drosophila screen for heat nociception identifies alpha2delta3 as an evolutionarily conserved pain gene. Cell 2010;143:628-638.
331. Huang F, Wang X, Ostertag EM, et al. TMEM16C facilitates Na(+)-activated K+ currents in rat sensory neurons and regulates pain processing. Nat Neurosci 2013;16:1284-1290.
332. Defazio G, Jankovic J, Giel JL, Papapetropoulos S. Descriptive Epidemiology of Cervical Dystonia. Tremor and Other Hyperkinetic Movements 2013;3:tre-03-193-4374-4372.
333. Charlesworth G, Plagnol V, Holmstrom KM, et al. Mutations in ANO3 Cause Dominant Craniocervical Dystonia: Ion Channel Implicated in Pathogenesis. American journal of human genetics 2012;91:1041-1050.

334. Fuchs T, Saunders-Pullman R, Masuho I, et al. Mutations in GNAL cause primary torsion dystonia. Nature Genetics 2013;45:88-92.
335. Santangelo G. Contributo clinico alla conoscenza delle forme familiari della dysbasia lordotica progressiva (spasmo di torsione). G Psychiatr Neuropathol 1934:52-77.
336. Gimenez-Roldan S, Delgado G, Marin M, Villanueva JA, Mateo D. Hereditary torsion dystonia in gypsies. Adv Neurol 1988;50:73-81.
337. Moretti P, Hedera P, Wald J, Fink J. Autosomal recessive primary generalized dystonia in two siblings from a consanguineous family. Movement disorders : official journal of the Movement Disorder Society 2005;20:245-247.
338. Chouery E, Kfoury J, Delague V, et al. A novel locus for autosomal recessive primary torsion dystonia (DYT17) maps to 20p11.22–q13.12. Neurogenetics 2008;9:287-293.
339. Fletcher NA. The genetics of idiopathic torsion dystonia. Journal of Medical Genetics 1990;27:409-412.
340. Suwanjang W, Holmstrom KM, Chetsawang B, Abramov AY. Glucocorticoids reduce intracellular calcium concentration and protects neurons against glutamate toxicity. Cell Calcium 2013;53:256-263.
341. Adra CN, Zhu S, Ko JL, et al. LAPTM5: a novel lysosomal-associated multispanning membrane protein preferentially expressed in hematopoietic cells. Genomics 1996;35:328-337.
342. Braunewell KH, Klein-Szanto AJ. Visinin-like proteins (VSNLs): interaction partners and emerging functions in signal transduction of a subfamily of neuronal Ca2+ -sensor proteins. Cell and tissue research 2009;335:301-316.
343. Gifford JL, Walsh MP, Vogel HJ. Structures and metal-ion-binding properties of the Ca2+-binding helix-loop-helix EF-hand motifs. The Biochemical journal 2007;405:199-221.
344. O'Callaghan DW, Tepikin AV, Burgoyne RD. Dynamics and calcium sensitivity of the Ca2+/myristoyl switch protein hippocalcin in living cells. The Journal of cell biology 2003;163:715-721.
345. Markova O, Fitzgerald D, Stepanyuk A, et al. Hippocalcin signaling via site-specific translocation in hippocampal neurons. Neuroscience letters 2008;442:152-157.
346. Zozulya S, Stryer L. Calcium-myristoyl protein switch. Proceedings of the National Academy of Sciences of the United States of America 1992;89:11569-11573.
347. Senin, II, Fischer T, Komolov KE, Zinchenko DV, Philippov PP, Koch KW. Ca2+-myristoyl switch in the neuronal calcium sensor recoverin requires different functions of Ca2+-binding sites. The Journal of biological chemistry 2002;277:50365-50372.
348. Ames JB, Lim S, Ikura M. Molecular structure and target recognition of neuronal calcium sensor proteins. Front Mol Neurosci 2012;5:10.

349. Weiergraber OH, Senin, II, Zernii EY, et al. Tuning of a neuronal calcium sensor. The Journal of biological chemistry 2006;281:37594-37602.
350. Mammen A, Simpson PJ, Nighorn A, et al. Hippocalcin in the olfactory epithelium: a mediator of second messenger signaling. Biochemical and biophysical research communications 2004;322:1131-1139.
351. Palmer CL, Lim W, Hastie PG, et al. Hippocalcin functions as a calcium sensor in hippocampal LTD. Neuron 2005;47:487-494.
352. Kerrigan TL, Daniel JW, Regan PL, Cho K. The role of neuronal calcium sensors in balancing synaptic plasticity and synaptic dysfunction. Front Mol Neurosci 2012;5:57.
353. Amici M, Doherty A, Jo J, et al. Neuronal calcium sensors and synaptic plasticity. Biochem Soc Trans 2009;37:1359-1363.
354. Tzingounis AV, Kobayashi M, Takamatsu K, Nicoll RA. Hippocalcin gates the calcium activation of the slow afterhyperpolarization in hippocampal pyramidal cells. Neuron 2007;53:487-493.
355. Kim KS, Kobayashi M, Takamatsu K, Tzingounis AV. Hippocalcin and KCNQ channels contribute to the kinetics of the slow afterhyperpolarization. Biophys J 2012;103:2446-2454.
356. Villalobos C, Andrade R. Visinin-like neuronal calcium sensor proteins regulate the slow calcium-activated afterhyperpolarizing current in the rat cerebral cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience 2010;30:14361-14365.
357. Kobayashi M, Masaki T, Hori K, et al. Hippocalcin-deficient mice display a defect in cAMP response element-binding protein activation associated with impaired spatial and associative memory. Neuroscience 2005;133:471-484.
358. Oh DY, Yon C, Oh KJ, Lee KS, Han JS. Hippocalcin increases phospholipase D2 expression through extracellular signal-regulated kinase activation and lysophosphatidic acid potentiates the hippocalcin-induced phospholipase D2 expression. J Cell Biochem 2006;97:1052-1065.
359. Oh DY, Cho JH, Park SY, et al. A novel role of hippocalcin in bFGF-induced neurite outgrowth of H19-7 cells. J Neurosci Res 2008;86:1557-1565.
360. Dang MT, Yokoi F, Cheetham CC, et al. An anticholinergic reverses motor control and corticostriatal LTD deficits in Dyt1 DeltaGAG knock-in mice. Behav Brain Res 2012;226:465-472.
361. Prescott IA, Dostrovsky JO, Moro E, Hodaie M, Lozano AM, Hutchison WD. Reduced paired pulse depression in the basal ganglia of dystonia patients. Neurobiology of Disease 2013;51:214-221.

362. Avchalumov Y, Volkmann CE, Ruckborn K, et al. Persistent changes of corticostriatal plasticity in dt mutant hamsters after age-dependent remission of dystonia. Neuroscience 2013;250C:60-69.
363. Iwabuchi S, Kakazu Y, Koh JY, Harata NC. Abnormal cytoplasmic calcium dynamics in central neurons of a dystonia mouse model. Neuroscience letters 2013;548:61-66.
364. Vemula SR, Puschmann A, Xiao J, et al. Role of Galpha(olf) in familial and sporadic adult-onset primary dystonia. Human Molecular Genetics 2013.
365. Schneider SA, Mohire MD, Trender-Gerhard I, et al. Familial dopa-responsive cervical dystonia. Neurology 2006;66:599-601.
366. Yang C, Pring M, Wear MA, et al. Mammalian CARMIL inhibits actin filament capping by capping protein. Dev Cell 2005;9:209-221.
367. Morgan H, Beck T, Blake A, et al. EuroPhenome: a repository for high-throughput mouse phenotyping data. Nucleic Acids Research 2010;38:D577-585.
368. Teraoka SN, Telatar M, Becker-Catania S, et al. Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. American journal of human genetics 1999;64:1617-1631.
369. Chun HH, Gatti RA. Ataxia-telangiectasia, an evolving phenotype. DNA Repair (Amst) 2004;3:1187-1196.
370. Subramony SH, Durr A. Ataxic disorders. Edinburgh: Elsevier, 2012.
371. Verhagen MM, Abdo WF, Willemsen MA, et al. Clinical spectrum of ataxia-telangiectasia in adulthood. Neurology 2009;73:430-437.
372. Saunders-Pullman R, Raymond D, Stoessl AJ, et al. Variant ataxia-telangiectasia presenting as primary-appearing dystonia in Canadian Mennonites. Neurology 2012;78:649-657.
373. Alterman N, Fattal-Valevski A, Moyal L, et al. Ataxia-telangiectasia: mild neurological presentation despite null ATM mutation and severe cellular phenotype. Am J Med Genet A 2007;143A:1827-1834.
374. Koepp M, Schelosky L, Cordes I, Cordes M, Poewe W. Dystonia in ataxia telangiectasia: report of a case with putaminal lesions and decreased striatal [123I]iodobenzamide binding. Movement disorders : official journal of the Movement Disorder Society 1994;9:455-459.
375. Agamanolis DP, Greenstein JI. Ataxia-telangiectasia. Report of a case with Lewy bodies and vascular abnormalities within cerebral tissue. J Neuropathol Exp Neurol 1979;38:475-489.
376. Eilam R, Peter Y, Groner Y, Segal M. Late degeneration of nigro-striatal neurons in ATM-/- mice. Neuroscience 2003;121:83-98.

377. Renwick A, Thompson D, Seal S, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics 2006;38:873-875.
378. Kline SL, Cheeseman IM, Hori T, Fukagawa T, Desai A. The human Mis12 complex is required for kinetochore assembly and proper chromosome segregation. J Cell Biol 2006;173:9-17.
379. Dephoure N, Zhou C, Villen J, et al. A quantitative atlas of mitotic phosphorylation. Proceedings of the National Academy of Sciences of the United States of America 2008;105:10762-10767.
380. Palmieri F. The mitochondrial transporter family SLC25: identification, properties and physiopathology. Mol Aspects Med 2013;34:465-484.
381. Palmieri F. The mitochondrial transporter family (SLC25): physiological and pathological implications. Pflugers Arch 2004;447:689-709.
382. Haitina T, Lindblom J, Renstrom T, Fredriksson R. Fourteen novel human members of mitochondrial solute carrier family 25 (SLC25) widely expressed in the central nervous system. Genomics 2006;88:779-790.
383. Mayr JA, Merkel O, Kohlwein SD, et al. Mitochondrial phosphate-carrier deficiency: a novel disorder of oxidative phosphorylation. American journal of human genetics 2007;80:478-484.
384. Palmieri L, Alberio S, Pisano I, et al. Complete loss-of-function of the heart/muscle-specific adenine nucleotide translocator is associated with mitochondrial myopathy and cardiomyopathy. Human Molecular Genetics 2005;14:3079-3088.
385. Wibom R, Lasorsa FM, Tohonen V, et al. AGC1 deficiency associated with global cerebral hypomyelination. N Engl J Med 2009;361:489-495.
386. Kobayashi K, Sinasac DS, Iijima M, et al. The gene mutated in adult-onset type II citrullinaemia encodes a putative mitochondrial carrier protein. Nature Genetics 1999;22:159-163.
387. Tamamori A, Okano Y, Ozaki H, et al. Neonatal intrahepatic cholestasis caused by citrin deficiency: severe hepatic dysfunction in an infant requiring liver transplantation. Eur J Pediatr 2002;161:609-613.
388. Camacho JA, Obie C, Biery B, et al. Hyperornithinaemia-hyperammonaemia-homocitrullinuria syndrome is caused by mutations in a gene encoding a mitochondrial ornithine transporter. Nature Genetics 1999;22:151-158.
389. Rosenberg MJ, Agarwala R, Bouffard G, et al. Mutant deoxynucleotide carrier is associated with congenital microcephaly. Nature Genetics 2002;32:175-179.

390. Spiegel R, Shaag A, Edvardson S, et al. SLC25A19 mutation as a cause of neuropathy and bilateral striatal necrosis. Annals of neurology 2009;66:419-424.
391. Huizing M, Iacobazzi V, Ijlst L, et al. Cloning of the human carnitine-acylcarnitine carrier cDNA and identification of the molecular defect in a patient. American journal of human genetics 1997;61:1239-1245.
392. Molinari F, Kaminska A, Fiermonte G, et al. Mutations in the mitochondrial glutamate carrier SLC25A22 in neonatal epileptic encephalopathy with suppression bursts. Clinical Genetics 2009;76:188-194.
393. Guernsey DL, Jiang H, Campagna DR, et al. Mutations in mitochondrial carrier family gene SLC25A38 cause nonsyndromic autosomal recessive congenital sideroblastic anemia. Nature Genetics 2009;41:651-653.
394. Silvestre JS, Tamarat R, Ebrahimian TG, et al. Vascular endothelial growth factor-B promotes in vivo angiogenesis. Circ Res 2003;93:114-123.
395. Hagberg CE, Falkevall A, Wang X, et al. Vascular endothelial growth factor B controls endothelial fatty acid uptake. Nature 2010;464:917-921.
396. Sun Y, Jin K, Childs JT, Xie L, Mao XO, Greenberg DA. Vascular endothelial growth factor-B (VEGFB) stimulates neurogenesis: evidence from knockout mice and growth factor administration. Dev Biol 2006;289:329-335.
397. Poesen K, Lambrechts D, Van Damme P, et al. Novel role for vascular endothelial growth factor (VEGF) receptor-1 and its ligand VEGF-B in motor neuron degeneration. The Journal of neuroscience : the official journal of the Society for Neuroscience 2008;28:10451-10459.
398. Li Y, Zhang F, Nagai N, et al. VEGF-B inhibits apoptosis via VEGFR-1-mediated suppression of the expression of BH3-only protein genes in mice and rats. J Clin Invest 2008;118:913-923.
399. Dhondt J, Peeraer E, Verheyen A, et al. Neuronal FLT1 receptor and its selective ligand VEGF-B protect against retrograde degeneration of sensory neurons. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 2011;25:1461-1473.
400. Du X, Wang Q, Hirohashi Y, Greene MI. DIPA, which can localize to the centrosome, associates with p78/MCRS1/MSP58 and acts as a repressor of gene transcription. Exp Mol Pathol 2006;81:184-190.
401. Bezy O, Elabd C, Cochet O, et al. Delta-interacting protein A, a new inhibitory partner of CCAAT/enhancer-binding protein beta, implicated in adipocyte differentiation. The Journal of biological chemistry 2005;280:11432-11438.

402. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105-132.
403. Straussberg R, Shorer Z, Weitz R, et al. Familial infantile bilateral striatal necrosis: clinical features and response to biotin treatment. Neurology 2002;59:983-989.
404. Thyagarajan D, Shanske S, Vazquez-Memije M, De Vivo D, DiMauro S. A novel mitochondrial ATPase 6 point mutation in familial bilateral striatal necrosis. Annals of neurology 1995;38:468-472.
405. Solano A, Roig M, Vives-Bauza C, et al. Bilateral striatal necrosis associated with a novel mutation in the mitochondrial ND6 gene. Annals of neurology 2003;54:527-530.
406. Kim IS, Ki CS, Park KJ. Pediatric-onset dystonia associated with bilateral striatal necrosis and G14459A mutation in a Korean family: a case report. J Korean Med Sci 2010;25:180-184.
407. Lal D, Becker K, Motameny S, et al. Homozygous missense mutation of NDUFV1 as the cause of infantile bilateral striatal necrosis. Neurogenetics 2013;14:85-87.
408. Basel-Vanagaite L, Muncher L, Straussberg R, et al. Mutated nup62 causes autosomal recessive infantile bilateral striatal necrosis. Annals of neurology 2006;60:214-222.
409. Spiegel R, Shaag A, Edvardson S, et al. SLC25A19 mutation as a cause of neuropathy and bilateral striatal necrosis. Ann Neurol 2009;66:419-424.
410. Kevelam SH, Rodenburg RJ, Wolf NI, et al. NUBPL mutations in patients with complex I deficiency and a distinct MRI pattern. Neurology 2013;80:1577-1583.
411. Sheftel AD, Stehling O, Pierik AJ, et al. Human ind1, an iron-sulfur cluster assembly factor for respiratory complex I. Molecular and cellular biology 2009;29:6059-6073.
412. Bych K, Kerscher S, Netz DJ, et al. The iron-sulphur protein Ind1 is required for effective complex I assembly. The EMBO Journal 2008;27:1736-1746.
413. Wolf NI, Seitz A, Harting I, et al. New pattern of brain MRI lesions in isolated complex I deficiency. Neuropediatrics 2003;34:156-159.
414. Tucker EJ, Mimaki M, Compton AG, McKenzie M, Ryan MT, Thorburn DR. Next-generation sequencing in molecular diagnosis: NUBPL mutations highlight the challenges of variant detection and interpretation. Human mutation 2012;33:411-418.
415. Wydro MM, Balk J. Reconstruction of human NUBPL p.D273QfsX31 in Yarrowia lipolytica provides insight into the pathogenic character of the NUBPL c.815-27T>C branch-site mutation which is associated with complex I deficiency. Dis Model Mech 2013.
416. Rosenberg MJ, Agarwala R, Bouffard G, et al. Mutant deoxynucleotide carrier is associated with congenital microcephaly. Nat Genet 2002;32:175-179.

417. Ortega-Recalde O, Fonseca DJ, Patino LC, et al. A novel familial case of diffuse leukodystrophy related to NDUFV1 compound heterozygous mutations. Mitochondrion 2013;13:749-754.
418. Schuelke M, Smeitink J, Mariman E, et al. Mutant NDUFV1 subunit of mitochondrial complex I causes leukodystrophy and myoclonic epilepsy. Nature Genetics 1999;21:260-261.
419. Marin SE, Mesterman R, Robinson B, Rodenburg RJ, Smeitink J, Tarnopolsky MA. Leigh syndrome associated with mitochondrial complex I deficiency due to novel mutations In NDUFV1 and NDUFS2. Gene 2013;516:162-167.
420. Sanz G, Schlegel C, Pernollet JC, Briand L. Comparison of odorant specificity of two human olfactory receptors from different phylogenetic classes and evidence for antagonism. Chem Senses 2005;30:69-80.
421. Garcia-Esparcia P, Schluter A, Carmona M, et al. Functional genomics reveals dysregulation of cortical olfactory receptors in Parkinson disease: novel putative chemoreceptors in the human brain. J Neuropathol Exp Neurol 2013;72:524-539.
422. Ansoleaga B, Garcia-Esparcia P, Llorens F, Moreno J, Aso E, Ferrer I. Dysregulation of brain olfactory and taste receptors in AD, PSP and CJD, and AD-related model. Neuroscience 2013.
423. Dreyer WJ. The area code hypothesis revisited: olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos. Proceedings of the National Academy of Sciences of the United States of America 1998;95:9072-9077.
424. Kang N, Koo J. Olfactory receptors in non-chemosensory tissues. BMB Rep 2012;45:612-622.
425. Griffin CA, Kafadar KA, Pavlath GK. MOR23 promotes muscle regeneration and regulates cell adhesion and migration. Dev Cell 2009;17:649-661.
426. Spehr M, Gisselmann G, Poplawski A, et al. Identification of a testicular odorant receptor mediating human sperm chemotaxis. Science 2003;299:2054-2058.
427. Spehr M, Schwane K, Heilmann S, Gisselmann G, Hummel T, Hatt H. Dual capacity of a human olfactory receptor. Current biology : CB 2004;14:R832-833.
428. Fukuda N, Yomogida K, Okabe M, Touhara K. Functional characterization of a mouse testicular olfactory receptor and its role in chemosensing and in regulation of sperm motility. Journal of Cell Science 2004;117:5835-5845.
429. Feinstein P, Mombaerts P. A contextual model for axonal sorting into glomeruli in the mouse olfactory system. Cell 2004;117:817-831.

430. Barnea G, O'Donnell S, Mancia F, et al. Odorant receptors on axon termini in the brain. Science 2004;304:1468.
431. Feinstein P, Bozza T, Rodriguez I, Vassalli A, Mombaerts P. Axon guidance of mouse olfactory sensory neurons by odorant receptors and the beta2 adrenergic receptor. Cell 2004;117:833-846.
432. El-Husseini AE, Fretier P, Vincent SR. Cloning and characterization of a gene (RNF22) encoding a novel brain expressed ring finger protein (BERP) that maps to human chromosome 11p15.5. Genomics 2001;71:363-367.
433. Yan Q, Sun W, Kujala P, Lotfi Y, Vida TA, Bean AJ. CART: an Hrs/actinin-4/BERP/myosin V protein complex required for efficient receptor recycling. Molecular biology of the cell 2005;16:2470-2482.
434. Cheung CC, Yang C, Berger T, et al. Identification of BERP (brain-expressed RING finger protein) as a p53 target gene that modulates seizure susceptibility through interacting with GABA(A) receptors. Proceedings of the National Academy of Sciences of the United States of America 2010;107:11883-11888.
435. Labonte D, Thies E, Kneussel M. The kinesin KIF21B participates in the cell surface delivery of gamma2 subunit-containing GABA receptors. Eur J Cell Biol 2014.
436. Hung AY, Sung CC, Brito IL, Sheng M. Degradation of postsynaptic scaffold GKAP and regulation of dendritic spine morphology by the TRIM3 ubiquitin ligase in rat hippocampal neurons. PLoS ONE 2010;5:e9842.
437. El-Husseini AE, Vincent SR. Cloning and characterization of a novel RING finger protein that interacts with class V myosins. The Journal of biological chemistry 1999;274:19771-19777.
438. Martins-de-Souza D, Gattaz WF, Schmitt A, et al. Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia. Eur Arch Psychiatry Clin Neurosci 2009;259:151-163.
439. Lorden JF, McKeon TW, Baker HJ, Cox N, Walkley SU. Characterization of the rat mutant dystonic (dt): a new animal model of dystonia musculorum deformans. The Journal of neuroscience : the official journal of the Society for Neuroscience 1984;4:1925-1932.
440. Aoyama T, Hata S, Nakao T, Tanigawa Y, Oka C, Kawaichi M. Cayman ataxia protein caytaxin is transported by kinesin along neurites through binding to kinesin light chains. J Cell Sci 2009;122:4177-4185.
441. Hayakawa Y, Itoh M, Yamada A, Mitsuda T, Nakagawa T. Expression and localization of Cayman ataxia-related protein, Caytaxin, is regulated in a developmental- and spatial-dependent manner. Brain Res 2007;1129:100-109.

442. Beales M, Lorden JF, Walz E, Oltmans GA. Quantitative autoradiography reveals selective changes in cerebellar GABA receptors of the rat mutant dystonic. The Journal of neuroscience : the official journal of the Society for Neuroscience 1990;10:1874-1885.
443. Lutes J, Lorden JF, Davis BJ, Oltmans GA. GABA levels and GAD immunoreactivity in the deep cerebellar nuclei of rats with altered olivo-cerebellar function. Brain Res Bull 1992;29:329-336.
444. Naudon L, Delfs JM, Clavel N, Lorden JF, Chesselet MF. Differential expression of glutamate decarboxylase messenger RNA in cerebellar Purkinje cells and deep cerebellar nuclei of the genetically dystonic rat. Neuroscience 1998;82:1087-1094.
445. Rock R, Schrauth S, Gessler M. Expression of mouse dchs1, fjx1, and fat-j suggests conservation of the planar cell polarity pathway identified in Drosophila. Dev Dyn 2005;234:747-755.
446. Ishiuchi T, Misaki K, Yonemura S, Takeichi M, Tanoue T. Mammalian Fat and Dachsous cadherins regulate apical membrane organization in the embryonic cerebral cortex. The Journal of cell biology 2009;185:959-967.
447. Mao Y, Mulvaney J, Zakaria S, et al. Characterization of a Dchs1 mutant mouse reveals requirements for Dchs1-Fat4 signaling during mammalian development. Development 2011;138:947-957.
448. UK Parkinson's Disease Consortium, Wellcome Trust Case Control Consortium 2, Spencer CCA, et al. Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. Hum Mol Genet 2011;20:345-353.
449. Satake W, Nakabayashi Y, Mizuta I, et al. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nature Genetics 2009;41:1303-1307.
450. Vandrovcova J, Pittman AM, Malzer E, et al. Association of MAPT haplotype-tagging SNPs with sporadic Parkinson's disease. Neurobiol Aging 2009;30:1477-1482.
451. Baker M, Litvan I, Houlden H, et al. Association of an extended haplotype in the tau gene with progressive supranuclear palsy. Human Molecular Genetics 1999;8:711-715.
452. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of Neurology, Neurosurgery & Psychiatry 1992;55:181.
453. Newman EJ, Breen K, Patterson J, Hadley DM, Grosset KA, Grosset DG. Accuracy of Parkinson's disease diagnosis in 610 general practice patients in the West of Scotland. Mov Disord 2009;24:2379-2385.

454. Hauw JJ, Daniel SE, Dickson D, et al. Preliminary NINDS neuropathologic criteria for Steele-Richardson-Olszewski syndrome (progressive supranuclear palsy). Neurology 1994;44:2015-2019.

455. Braak H, Del Tredici K, Rüb U, de Vos RAI, Jansen Steur ENH, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging 2003;24:197-211.

456. Ince PG, Clark B, Holton JL, Revesz T, Wharton S. Disorders of Movement and System Degenerations. In: Love S, Louis DN, Ellison DW, eds. Greenfield's Neuropathology, 8th ed. London: Arnold, 2008: 889-1030.

457. Myers AJ, Pittman AM, Zhao AS, et al. The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. Neurobiol Dis 2007;25:561-570.

458. Higashi S, Iseki E, Yamamoto R, et al. Concurrence of TDP-43, tau and alpha-synuclein pathology in brains of Alzheimer's disease and dementia with Lewy bodies. Brain Res 2007;1184:284-294.

459. Goris A, Williams-Gray CH, Clark GR, et al. Tau and α-synuclein in susceptibility to, and dementia in, Parkinson's disease. Ann Neurol 2007;62:145-153.

460. Williams-Gray CH, Evans JR, Goris A, et al. The distinct cognitive syndromes of Parkinson's disease: 5 year follow-up of the CamPaIGN cohort. Brain 2009;132:2958-2969.

461. Jellinger KA. Neuropathological Aspects of Alzheimer Disease, Parkinson Disease and Frontotemporal Dementia. Neurodegenerative Dis 2008;5:118-121.

462. Naj AC, Jun G, Beecham GW, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 2011;43:436-441.

463. Devine MJ, Lewis PA. Emerging pathways in genetic Parkinson's disease: tangles, Lewy bodies and LRRK2. FEBS J 2008;275:5748-5757.

464. Haggerty T, Credle J, Rodriguez O, et al. Hyperphosphorylated Tau in an α-synuclein-overexpressing transgenic model of Parkinson's disease. The European journal of neuroscience 2011.

465. Wills J, Jones J, Haggerty T, Duka V, Joyce JN, Sidhu A. Elevated tauopathy and alpha-synuclein pathology in postmortem Parkinson's disease brains with and without dementia. Experimental Neurology 2010;225:210-218.

466. Zimprich A, Benet-Pagès A, Struhal W, et al. A Mutation in VPS35, Encoding a Subunit of the Retromer Complex, Causes Late-Onset Parkinson Disease. The American Journal of Human Genetics 2011;89:168-175.

467. Doty RL, Shaman P, Dann M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. Physiol Behav 1984;32:489-502.

468. Silveira-Moriyama L, Petrie A, Williams DR, et al. The use of a color coded probability scale to interpret smell tests in suspected parkinsonism. Mov Disord 2009;24:1144-1153.

469. Alcalay RN, Siderowf A, Ottman R, et al. Olfaction in Parkin heterozygotes and compound heterozygotes: the CORE-PD study. Neurology 2011;76:319-326.

470. Khan NL, Katzenschlager R, Watt H, et al. Olfaction differentiates parkin disease from early-onset parkinsonism and Parkinson disease. Neurology 2004;62:1224-1226.

471. Healy DG, Falchi M, O'Sullivan SS, et al. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol 2008;7:583-590.

472. Saunders-Pullman R, Stanley K, Wang C, et al. Olfactory dysfunction in LRRK2 G2019S mutation carriers. Neurology 2011;77:319-324.

473. Silveira-Moriyama L, Guedes LC, Kingsbury A, et al. Hyposmia in G2019S LRRK2-related parkinsonism: clinical and pathologic data. Neurology 2008;71:1021-1026.

474. Silveira-Moriyama L, Munhoz RP, de J Carvalho M, et al. Olfactory heterogeneity in LRRK2 related Parkinsonism. Mov Disord 2010;25:2879-2883.

475. Chartier-Harlin MC, Dachsel JC, Vilarino-Guell C, et al. Translation initiator EIF4G1 mutations in familial Parkinson disease. American journal of human genetics 2011;89:398-406.

476. Khan NL, Jain S, Lynch JM, et al. Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data. Brain : a journal of neurology 2005;128:2786-2796.

477. Sheerin UM, Charlesworth G, Bras J, et al. Screening for VPS35 mutations in Parkinson's disease. Neurobiology of aging 2012;33:838 e831-835.

478. Cann HM, de Toma C, Cazes L, et al. A human genome diversity cell line panel. Science 2002;296:261-262.

479. Guerreiro RJ, Washecka N, Hardy J, Singleton A. A thorough assessment of benign genetic variability in GRN and MAPT. Human mutation 2010;31:E1126-1140.

480. Zhang X. Exome sequencing greatly expedites the progressive research of Mendelian diseases. Front Med 2014;8:42-57.

481. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461:272-276.

482. Haack TB, Hogarth P, Kruer MC, et al. Exome sequencing reveals de novo WDR45 mutations causing a phenotypically distinct, X-linked dominant form of NBIA. American journal of human genetics 2012;91:1144-1149.
483. Picollo A, Malvezzi M, Accardi A. TMEM16 Proteins: Unknown Structure and Confusing Functions. J Mol Biol 2015;427:94-105.
484. Vilariño-Güell C, Wider C, Ross Owen A, et al. VPS35 Mutations in Parkinson Disease. The American Journal of Human Genetics 2011;89:162-167.
485. Charlesworth G, Bhatia KP, Wood NW. No pathogenic GNAL mutations in 192 sporadic and familial cases of cervical dystonia. Mov Disord 2014;29:154-155.
486. Erro R, Bhatia KP, Hardy J. GNAL mutations and dystonia. JAMA Neurol 2014;71:1052-1053.
487. Huttenlocher J, Kruger R, Capetian P, et al. EIF4G1 is neither a strong nor a common risk factor for Parkinson's disease: evidence from large European cohorts. J Med Genet 2015;52:37-41.
488. Pennisi E. Genomics. 1000 Genomes Project gives new map of genetic diversity. Science 2010;330:574-575.
489. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev 2003;17:419-437.
490. Caceres JF, Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends in genetics : TIG 2002;18:186-193.
491. Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett 2005;579:1900-1903.
492. Bisio A, Nasti S, Jordan JJ, et al. Functional analysis of CDKN2A/p16INK4a 5'-UTR variants predisposing to melanoma. Human Molecular Genetics 2010;19:1479-1491.
493. Abelson JF, Kwan KY, O'Roak BJ, et al. Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 2005;310:317-320.
494. Bonafe L, Dermitzakis ET, Unger S, et al. Evolutionary comparison provides evidence for pathogenicity of RMRP mutations. PLoS Genetics 2005;1:e47.
495. Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell 2009;136:777-793.
496. Martin MP, Dean M, Smith MW, et al. Genetic acceleration of AIDS progression by a promoter variant of CCR5. Science 1998;282:1907-1911.
497. Bray NJ, Jehu L, Moskvina V, et al. Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. Human Molecular Genetics 2004;13:2885-2892.
498. Exner M, Minar E, Wagner O, Schillinger M. The role of heme oxygenase-1 promoter polymorphisms in human disease. Free Radic Biol Med 2004;37:1097-1104.

499. Lettice LA, Heaney SJ, Purdie LA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics 2003;12:1725-1735.

500. Duan J, Wainwright MS, Comeron JM, et al. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Human Molecular Genetics 2003;12:205-216.

501. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012;13:762-775.

502. Raska P, Zhu X. Rare variant density across the genome and across populations. BMC Proc 2011;5 Suppl 9:S39.

503. Keen-Kim D, Mathews CA, Reus VI, et al. Overrepresentation of rare variants in a specific ethnic group may confuse interpretation of association analyses. Human Molecular Genetics 2006;15:3324-3328.